

Introduction To Statistics

Instructor: Dr. Henry Hsu

Prepared By: Fairouz Makhoul

TASC/Advanced Support Team
Center for Information Technology
National Institute of Health

Contents:

- Definitions
- Describing Data
- Sampling distribution of the sample mean
- Statistical Inference
- Correlation and Regression

Definition

Statistics consist of a set of methods and rules for organizing and interpreting observations.

These statistical procedures help ensure that the data are presented and interpreted in an accurate and informative way.

Populations AND Samples

A *population* is the entire group of individuals that a researcher wishes to study. **Entire group** means every single individual.

A *sample* is a set of individuals selected from a population, usually intended to represent the population in a study.

Parameters AND Statistics

A *parameter* is a value, usually numeric that describes a population. It may be obtained from a single measurements or it may be derived from a set of measurements from the population

A *statistic* is a value, usually numeric, that describes a sample. A statistic may be obtained from a single measurement or it may be derived from from a set of measurements from the sample.

NOTE: Typically every *population parameter* has a corresponding *sample statistic*.

Descriptive AND Inferential Statistical Methods

Descriptive Statistical Methods

They are methods that summarize, organize, and simplify data. In other words, they are techniques that takes raw scores and summarize them in a form that is more manageable. The average, median, standard deviation are some examples of descriptive Statistics.

Inferential statistics Methods

They are techniques that allows us to study samples and then make a generalizations about the populations from which they are selected.

Sampling Error

It is the amount of error that exists between a sample statistics and a corresponding population parameter.

Describing Data Graphically

- Histograms
- Frequency Polygons
- Bar Graph
- Pie Charts
- Stem and leaf Plots

Histogram

- It is a visual illustration of the frequency distribution.
- It is used when the data are measured on an interval or ratio scale.
- It shows the shape of the data in a graphical form.
- It is a good tool for differentiating the frequencies of class intervals.
- To construct a Histogram, vertical bars are drawn above each score so that
 - The height of the bar corresponds to the frequency
 - The width of the bar extends to the real limits of the scores.

Example 1:

Sketch a frequency distribution histogram for each set of data given in the following tables. Describe the shape of the distribution.

Table I

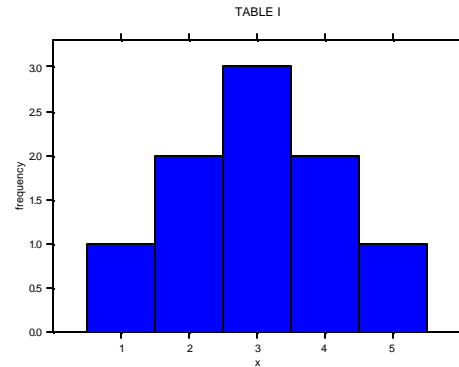
x	f r e q u e n c y
5	1
4	2
3	3
2	2
1	1

Table II

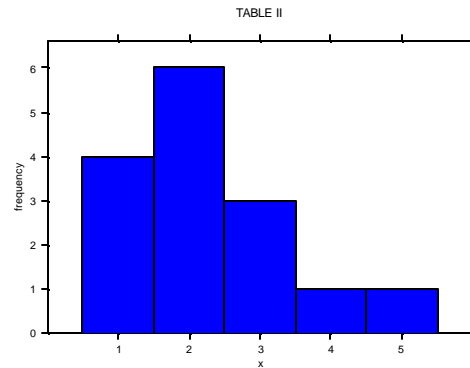
x	f r e q u e n c y
5	1
4	1
3	3
2	6
1	4

Table III

x	f r e q u e n c y
5	4
4	6
3	3
2	1
1	1

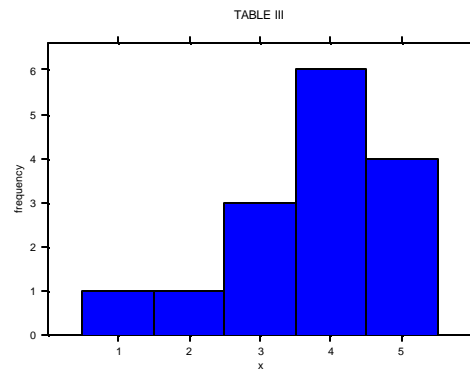


Symmetric distribution



Skewed to the right

Positively Skewed



Skewed to the left

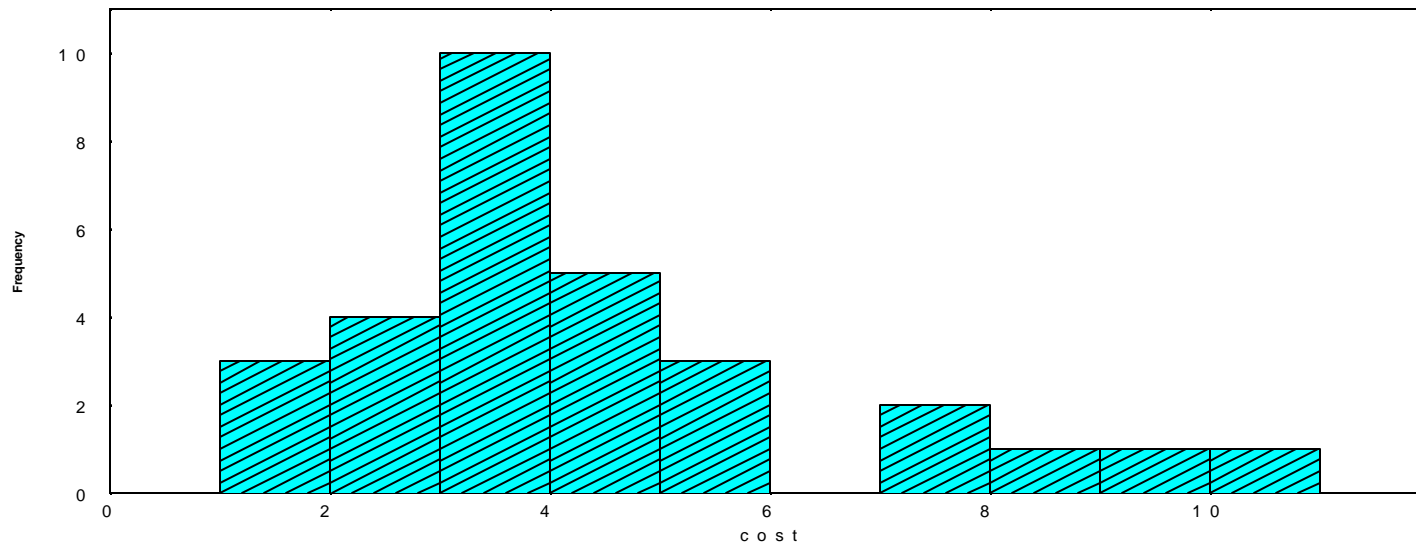
Negatively Skewed

Example 2:

Data Set

It represent the cost of a sample of 30 parcel-post mailing by a company in dollars:

3.67 2.75 5.47 4.65 3.32 2.09
1.83 10.94 1.93 3.89 7.20 2.78
3.34 7.80 3.20 3.21 3.55 3.53
3.64 4.95 5.42 8.64 4.84 4.10
9.15 3.45 5.11 1.97 2.84 4.15



Looking at the data set we have we can not reveal much about the data set. But looking at the histogram we can see the following:

- No parcel cost between \$6 and \$ 7 dollars.
- The parcels that costs between \$3 and \$4 have the highest frequency.
- Not many parcels cost more than \$7.

Frequency Distribution Polygon

It is intended for use with intervals or ratio scale.

To construct a frequency distribution polygon, a single dot is drawn above each score so that

- The dot is centered above the score
- The height of the dot corresponds to the frequency

A continuous line is then drawn connecting these dots. The graph is completed by drawing a line down to the X-axis at each end of the range of scores.

Example 3:

Draw a frequency polygon for the data in the tables in example 1.

Table I

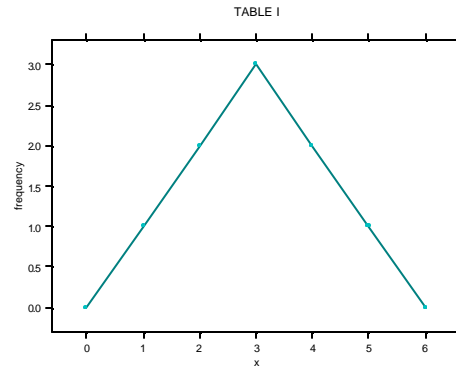


Table II

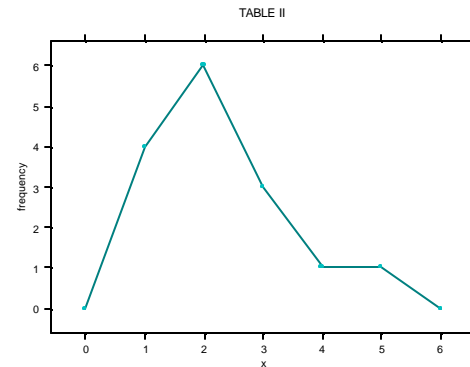
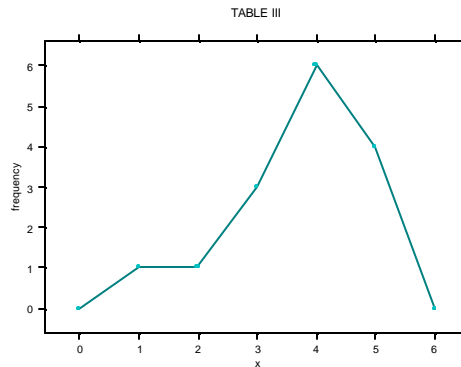


Table III



Bar Graph

- A bar graph is used when the data is measured in a nominal or ordinal scale.
- To construct a bar graph
 - A vertical bar is drawn above each score (or category) so that
 - The height of the bar corresponds to the frequency or percentage.
 - There is a space separating each bar from the next

Example 4:

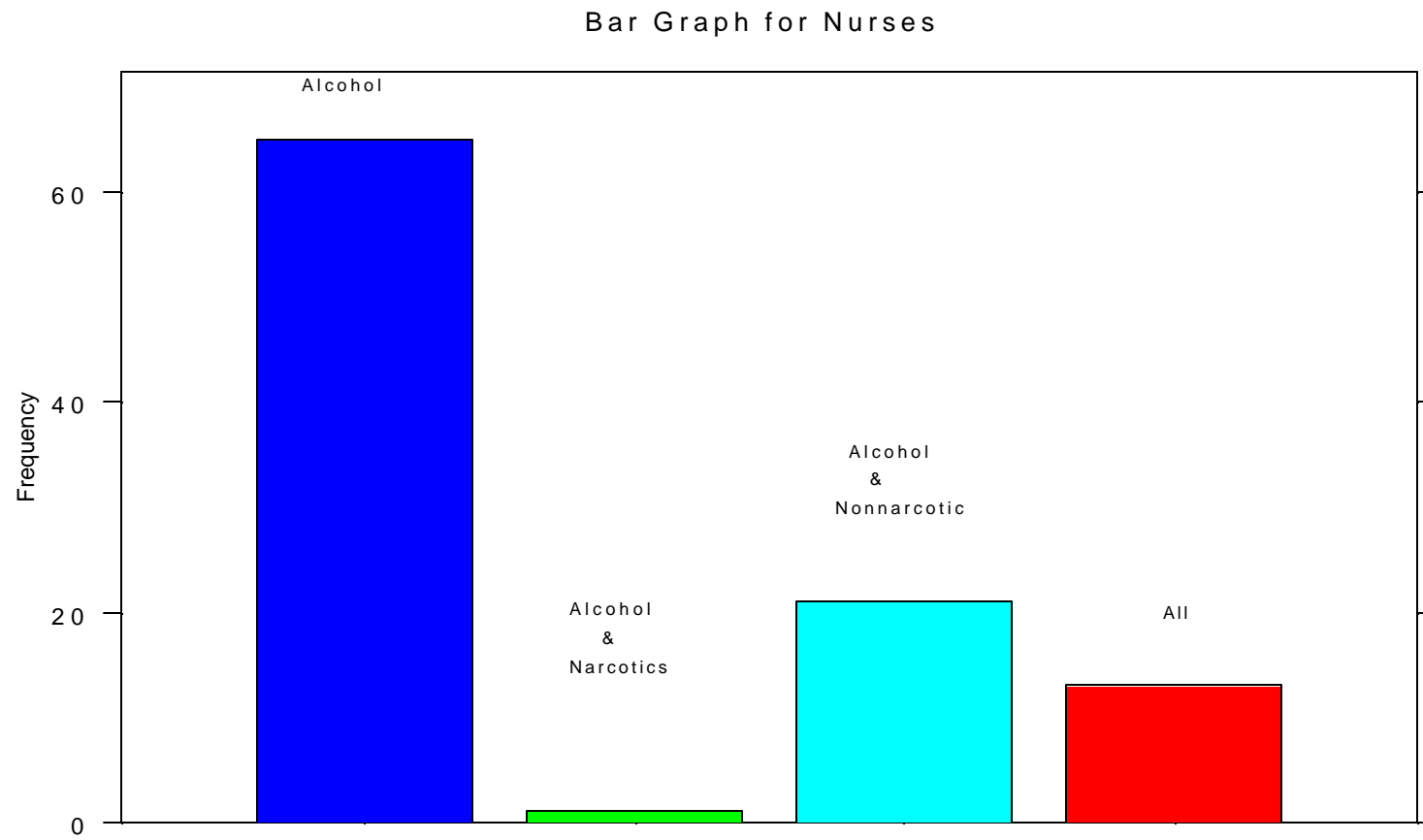
How prevalent is alcoholism among medical professionals? Does ease access to drug tend to encourage their use? Is there a difference between the rates of addiction for nurses and physicians? In an attempt to answer these and other questions, researchers conducted a survey of nurses and physicians who considered themselves alcoholics, were members of Alcoholics Anonymous, and had been completely abstinent for at least 1 calendar year immediately prior to being interviewed.

One aspect of the study concerned the subjects' addiction to other drugs. Subjects were asked if they had used a drug outside the hospital setting and, if so, had they been addicted to it. The results shown in the table were observed

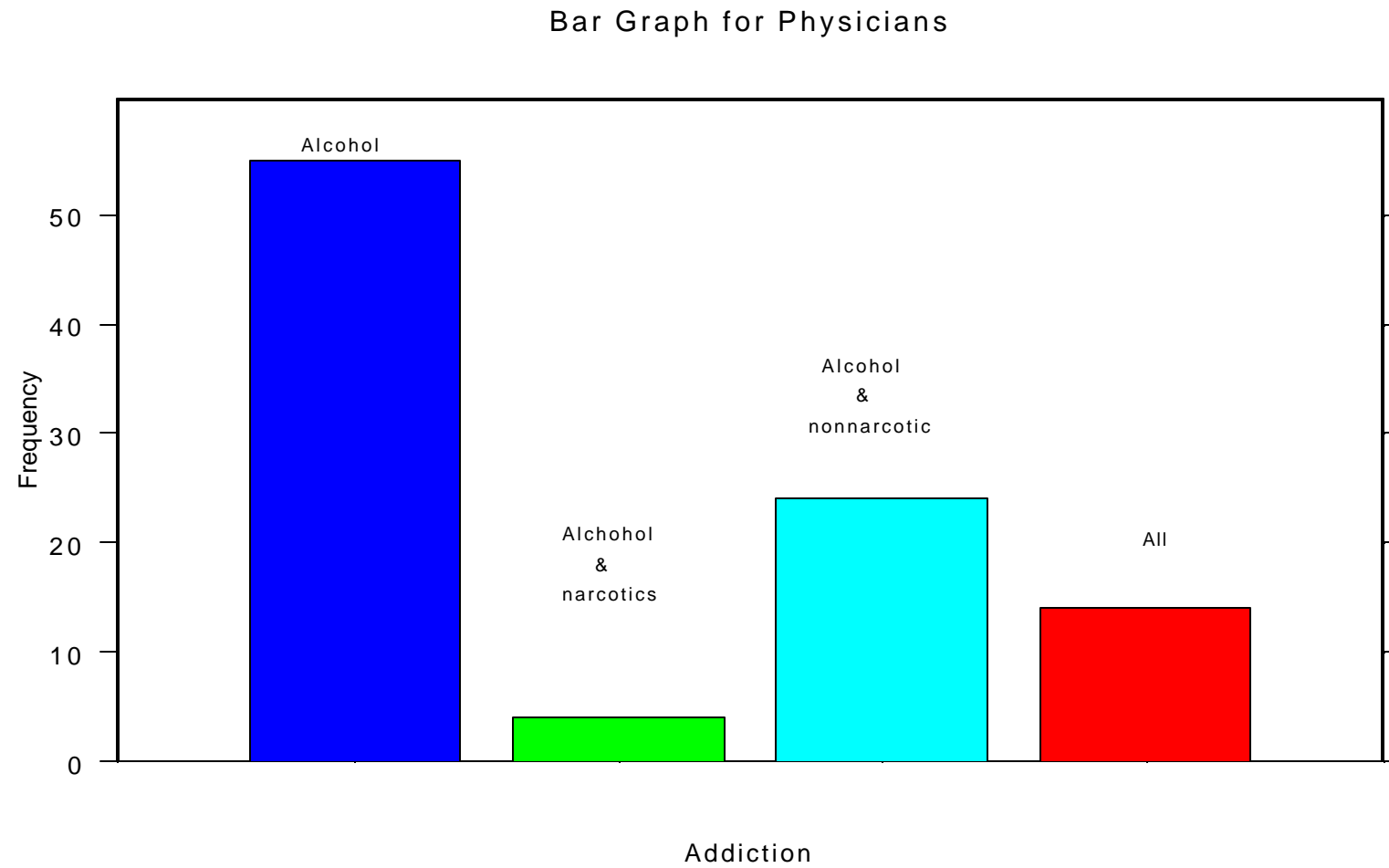
Addiction	Number of Nurses	Number of Physicians
Alcohol Only	65	55
Both alcohol and narcotics	1	4
Both alcohol and nonnarcotic drugs	21	24
Alcohol, narcotics, and nonnarcotic drugs	13	14
Totals	100	97

- a.** Construct a bar graph to describe addiction among the nurses interviewed.
- b.** Construct a bar graph to describe addiction among the nurses interviewed.
- c.** Compare the two figures you constructed in parts a and b. Does there appear to be a difference between the rates of addiction for the two groups of subjects? Explain.

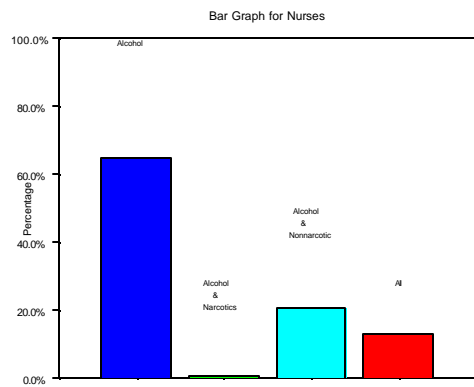
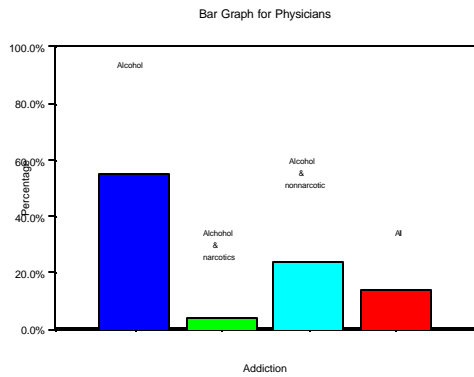
a.



b.



C.



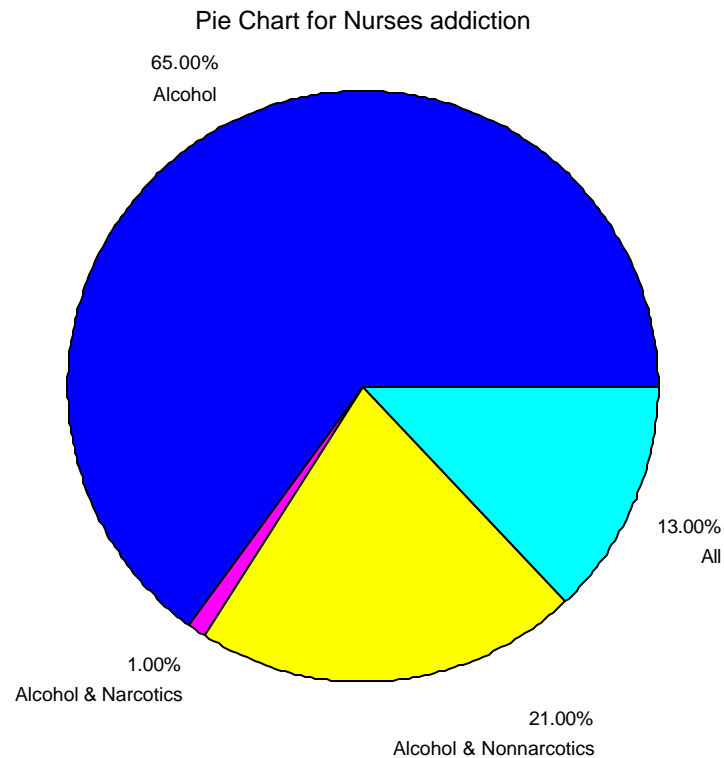
It appears that the rate of alcohol addiction for the nurses group is higher than the physicians group.

Pie Charts

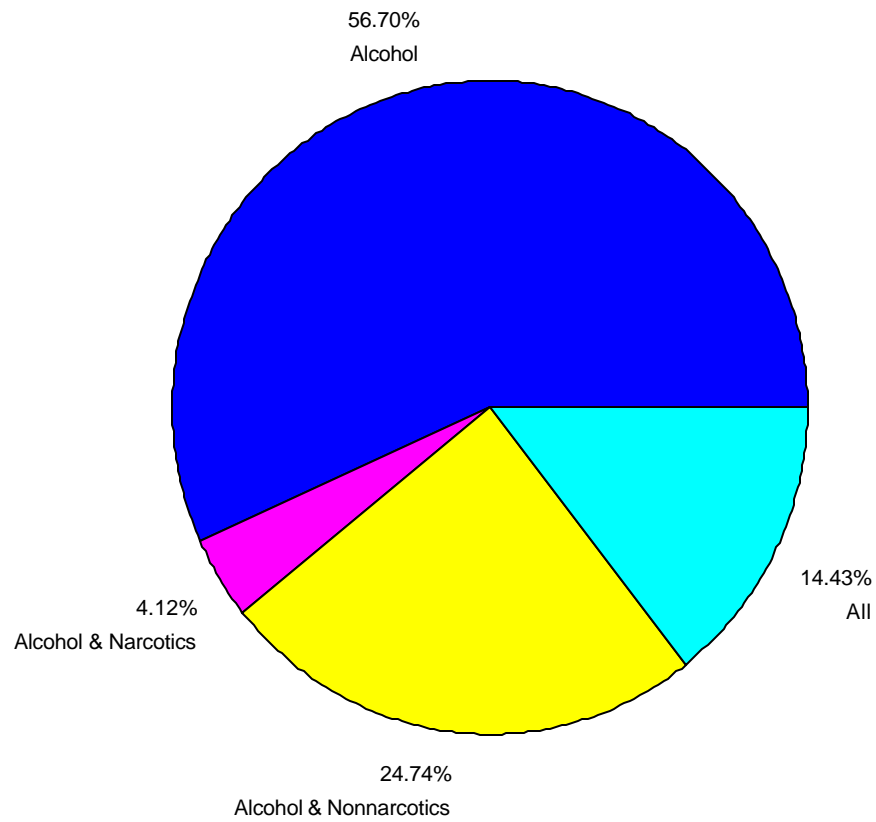
- The Pie chart is a commonly used graphical device for presenting qualitative data.
- To construct a pie chart
 - draw a circle
 - use relative frequency to subdivide the circle into sectors or parts, that correspond to the relative frequency for each class

Example 5:

Answer part a and b in Example 4 using a pie chart instead of a bar graph.



Pie Chart for Physicans Addiction



Stem & Leaf Plot

It requires that each score be separated into two parts

- The first digit (or digits) called the *Stem*.
- The last digit or digits called the *Leaf*.

It provides both a listing of the scores and a picture of the distribution.

Example 6:

Construct a stem and leaf plot for the following data:

3.6 2.7 5.4 4.6 3.3 2.0 3.9

1.8 1.9 7.2 2.8 3.3 7.8 4.1

3.2 3.2 3.5 3.5 3.7 4.9 2.1

5.5 5.1 8.6 4.1 2.8 3.4 3.9

Decimal point is at the colon

1 : 89

2 : 01788

3 : 22334556799

4 : 1169

5 : 145

6 :

7 : 28

8 : 6

Exercises

1. Sketch a histogram and a polygon showing the distribution of scores presented in the following table. What is the shape of the distribution.

<i>X</i>	<i>f r e q u e n c y</i>
1 5	2
1 4	5
1 3	6
1 2	3
1 1	2
1 0	1

2. The following data are attitude scores for a sample of 25 students. A high score indicates a positive attitude, and a low score indicate a negative attitude.

9	73	62	52	14
31	26	74	61	13
79	58	16	62	7
77	9	30	18	23
42	78	10	66	82

a. Construct a stem and leaf display to organize these data.

b. Looking at the distribution of scores, which of the following descriptions best fit these data?

1. This group has generally positive attitude.
2. This group has a generally negative attitude.
3. This group is sharply split with attitudes at both extremes.

3. Under what circumstances should you use a bar graph instead of a histogram to display a frequency distribution?

4. The following data are quiz scores from two different sections

<u>SECTION I</u>			<u>SECTION II</u>		
9	6	8	4	7	8
10	8	3	6	3	7
7	8	8	4	6	5
7	5	10	10	3	6
9	6	7	7	4	6

- Organize the scores from each section in a frequency distribution histogram.
- Describe the general difference between the two distribution

Describing Data Numerically

Descriptive Statistics

- Objectives:
 - Distinguish between measures of location, measures of variability and measures of shape.
 - Compute and understand the mean, median, mode, range, variance, standard deviation.
 - Differentiate between sample and population variance and standard deviation.

Measures of Location (Central Tendency)

- It is a statistical measure that identifies a single score as a representative for the entire distribution.
- The goal of central tendency is to find the single score that is most typical or most representative of the entire group.
- There are three different ways to measure the central tendency: the mode, median and the mean.

- * **Mode:** the most frequently occurring value in a set of data.
- * **Median:** is the middle value in an ordered array of numbers.
- * **Mean:** is the average of the values and is computed by summing the values and dividing by their number.

- As we can see that the above measures are computed differently and have different characteristics.

- To decide which of the three measures is best for any particular distribution, we have to keep in mind that the general purpose of central tendency is to find the single most representative score.

M o d e

- It is the most frequently occurring value in a set of data.
- If there is a tie in the most frequently value, this means we have more than one mode.
- It is used when preference study is done for different groups.

Example 7 :

Determine the mode for the following numbers.

2	4	8	4	6	2	7
8	4	3	8	9	4	3

To locate the mode it is helpful to sort the data

2	2	3	3	4	4	4
4	6	7	8	8	8	9

So 4 is the mode.

Median

- The median is the score value that has half of the data below and half of the data above.
- To compute the median of a set of data that has n observations we do the following:
 - Arrange the observations in an ascending order.
 - If the number of observations is odd. Then the median will be the middle observation which is the $\frac{n+1}{2}$ observation.
 - If the number of observations is even. Then the mean will be the average of the two middle observations.

Example 8:

- Compute the median for the following numbers:

16 28 29 13 17 20 11 34

32 27 25 30 19 18 33

- Arrange the data in an ascending order.

11 13 16 17 18 19 20 25

27 28 29 30 32 33 34

- Since the number of observation is odd. The median is the middle value which can be located by

$$\frac{n + 1}{2} = \frac{15 + 1}{2} = 8$$

- So, the median is 25.

Example 9:

- Compute the median for the following numbers:

213 345 609 73 167

243 444 524 199 682

- Arrange the data in an ascending order.

73 167 199 213 243

345 444 524 609 682

- Since the number of observations is even. The median is the average of the middle two value i.e.

$$\text{median} = \frac{243 + 345}{2} = 294$$

M e a n

- The mean for a distribution is the sum of the scores divided by the number of scores.
- It is commonly known as the arithmetic average.
- The mean for a population is denoted by **m**.
- The mean of a sample is denoted by \bar{x} .

Example 10:

Compute the mean for the following sample scores

6, 1, 8, 0, 5

$$\bar{x} = \frac{\sum x}{n} = \frac{20}{5} = 4$$

Selecting a measure of Tendency

When to use the mean?

Generally the mean is considered the best of the three measures of central tendency. But there is situation where it is either impossible to compute the mean or the mean is not particularly representative.

When to use the mode?

The mode can be used with any scale of measurements which makes it very flexible.

For the nominal scale, it is impossible or meaningless to calculate the mean or the median, so the mode is the only way to describe central tendency.

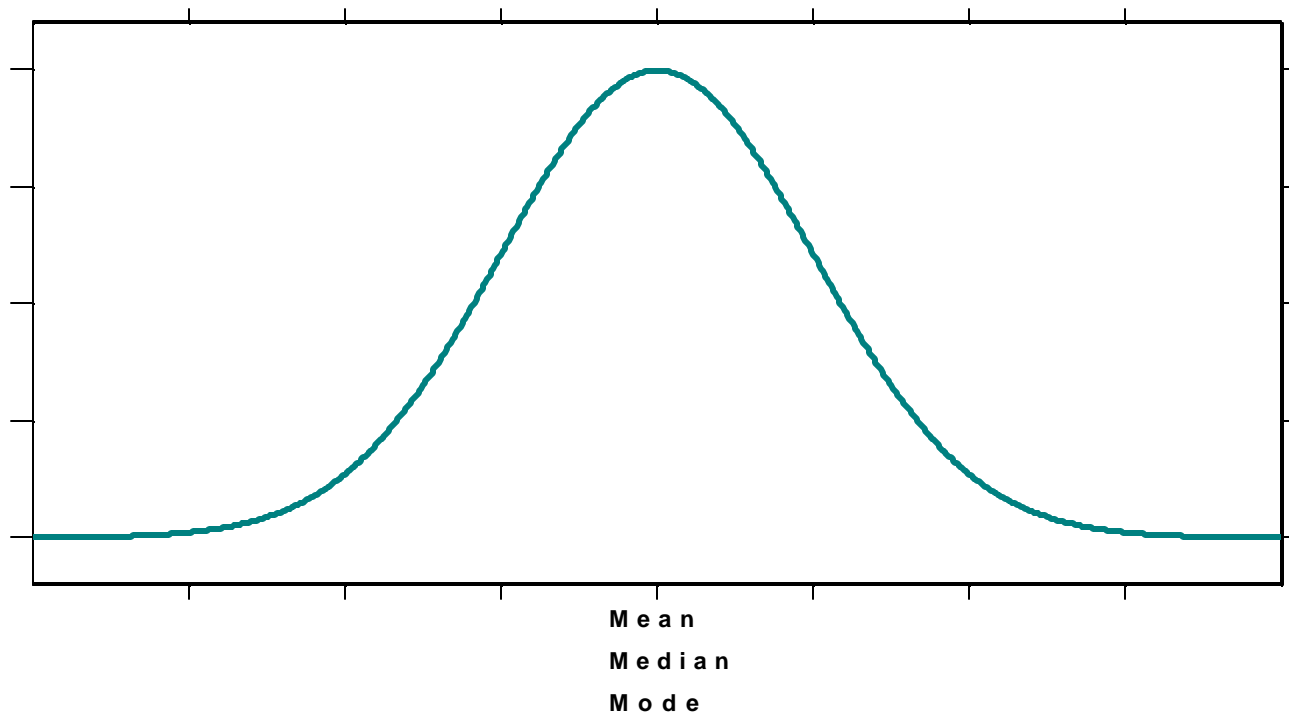
When to use the median?

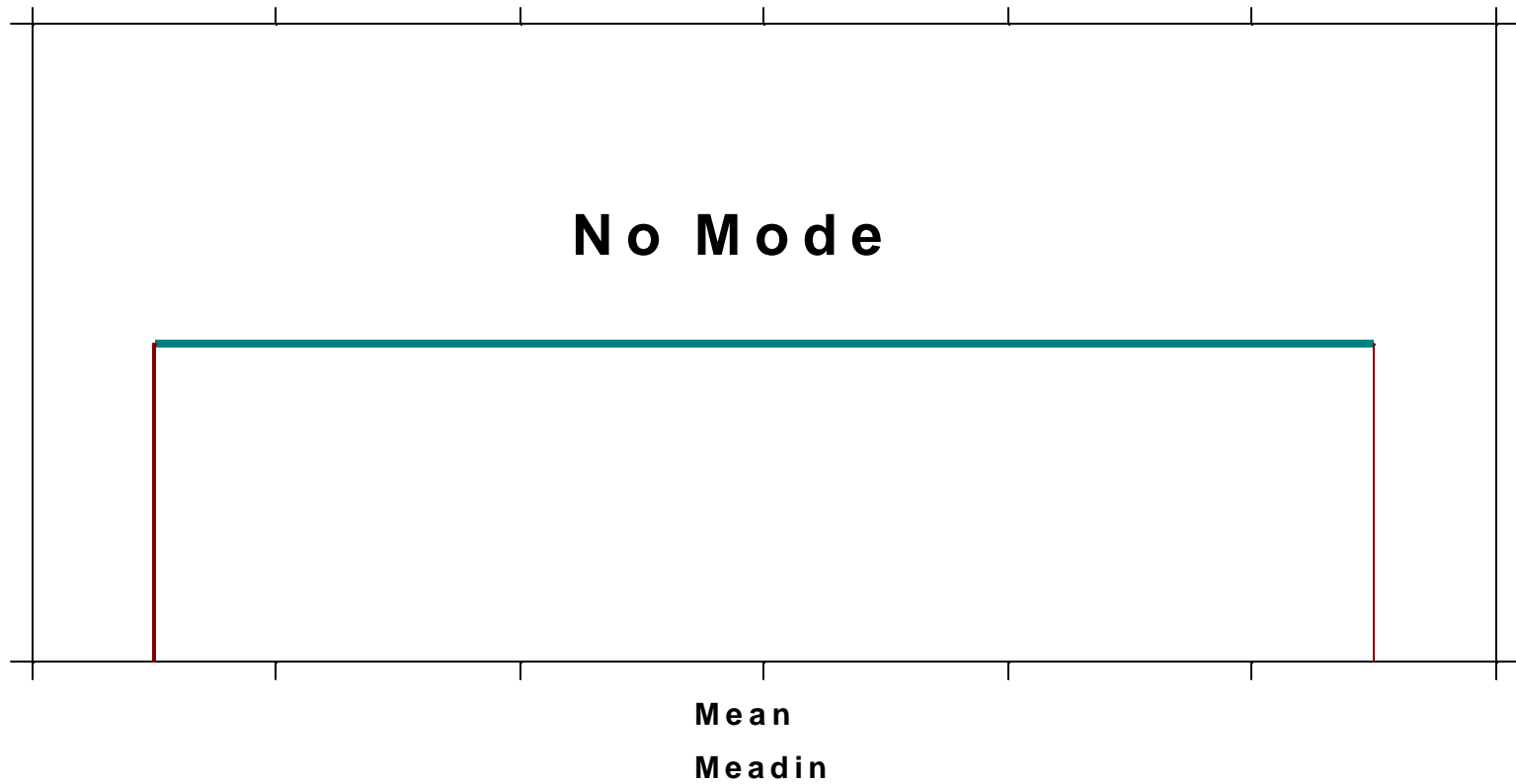
There are four situation where the median serves as an alternative to the mean.

- There are a few extreme scores in the distribution
- Some scores have undetermined values
- There is an open-ended distribution
- The data are measured on an ordinal scale.

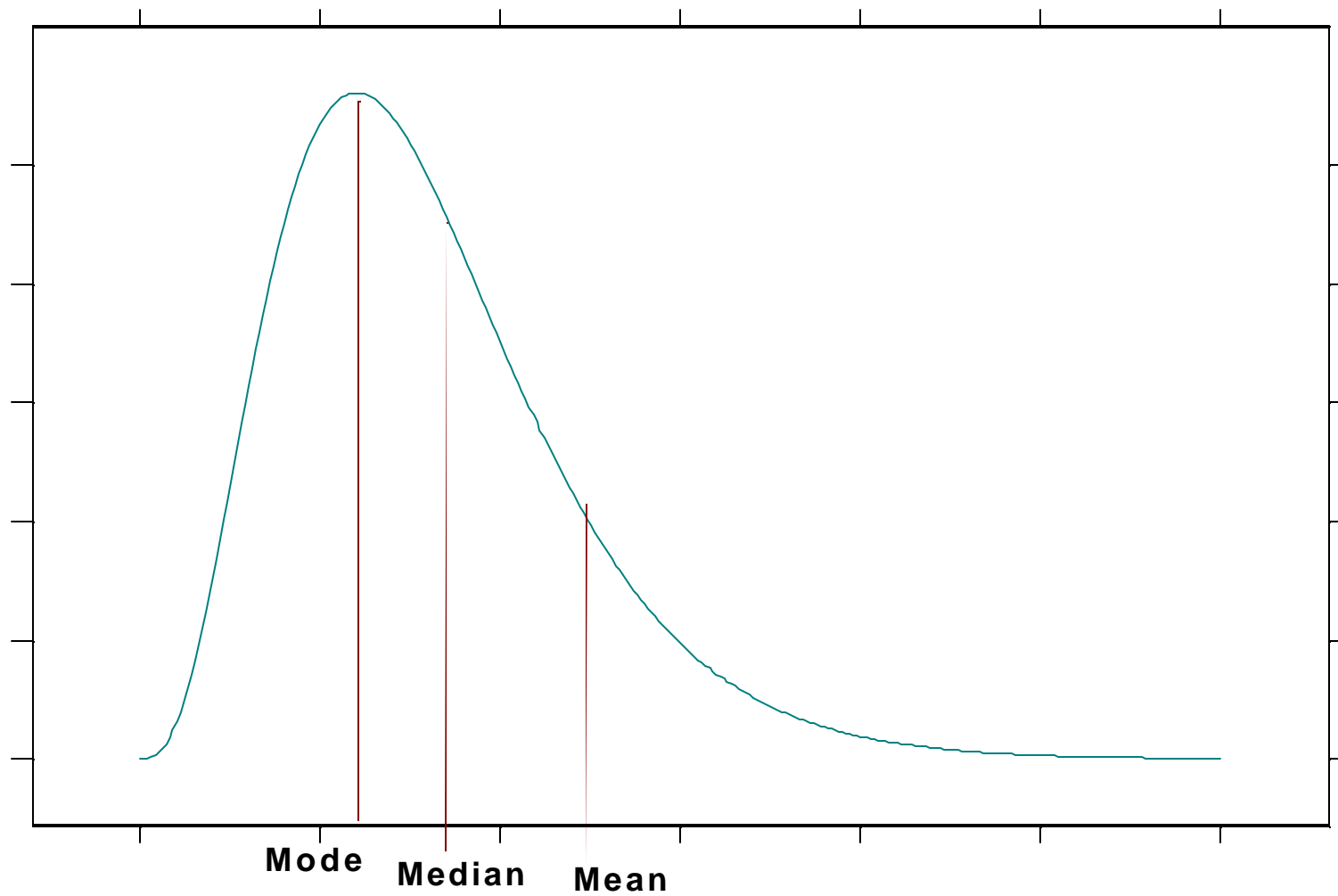
Central Tendency And the Shape of the distribution

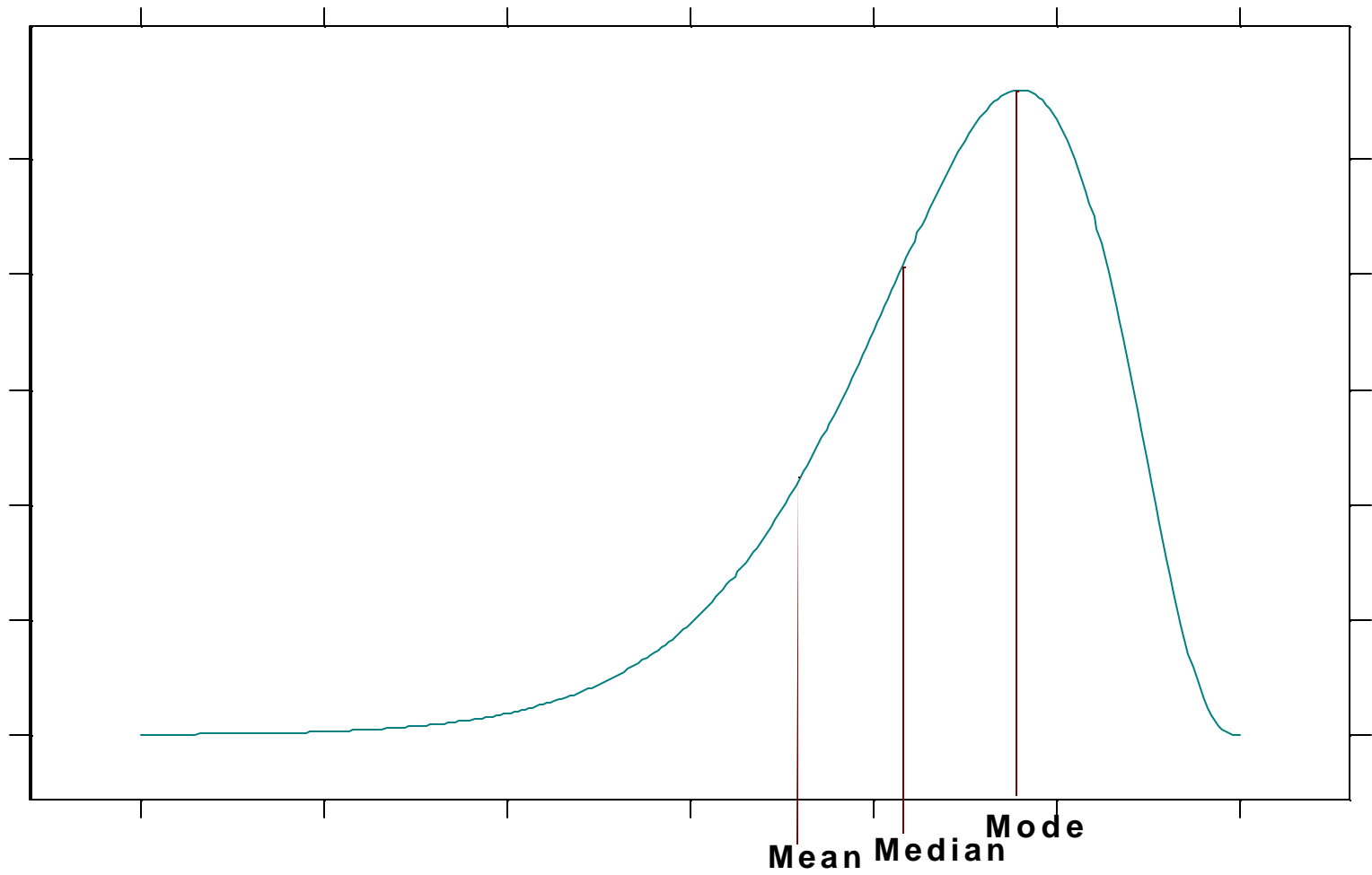
Symmetric distributions





Skewed Distribution





Exercises

5. Explain why the mean is not a necessarily a good measure for central tendency for a skewed distribution.
6. Explain why the mean and the median are probably not a good measure of central tendency for a symmetric, bimodal distribution.
7. Under what circumstances will the mean, the median, and the mode all have the same value?

8. A psychologist would like to determine how many errors are, made on average, before a rat can learn a particular maze. A sample of $n=10$ rats is obtained, and each rat is tested on the maze. The scores for the first 9 rats are as follows:

6, 2, 4, 5, 3, 7, 6, 2, 1.

a. Calculate the mean and the median for these data. On the average, how many errors does each rat make?

b. The tenth rat in the sample committed 100 errors before mastering the maze. When this rat is included in the sample, what happens to the mean? What happens to the median? What general conclusion can be drawn from this result.

9. Under what circumstances is the mode the preferred measure of central tendency.

10. If you change the value of a single score in a distribution, you will sometimes change the median, and sometimes leave the median unaffected. Describe the circumstances where the median would change and where the median would not change.

11. A distribution of a set of scores has a mean of 71 and a median of 79. Is it more likely that the distribution is symmetrical, positively skewed, or negatively skewed?

Variability

- It provides a quantitative measure of the degree to which scores in a distribution are spread out or clustered together.
- A good measure of variability should provide an accurate picture of the spread of the distribution.
- It also, give an indication of how well an individual score (or group of scores) represent the entire population.

- Here we will consider three different measures of variability

- * Range

- * Variance and standard deviation

The Range

It is the distance between the largest score and the smallest one in a distribution i.e.

$$\text{Range} = \text{Max} - \text{Min}$$

Example 11:

Find the range for the following scores

3, 8, 9, 10, 12 , 25, 4

$$\text{Range} = 25 - 3 = 22$$

The problem with the range it is determined completely by the smallest and the largest values and ignores the rest values in the distribution. For this reason the range is considered to be a crude and unreliable measure of variability.

Example 12:

Consider the following two sets of scores (assume the data is the scores of a quiz for two sections):

distribution I

1, 8, 9, 9, 10, 10

(almost all of the students mastered the material)

$$\text{Range} = 10 - 1 = 9$$

distribution II

1, 2, 4, 6, 8, 10

(There is a wide range of abilities)

$$\text{Range} = 10 - 1 = 9$$

As we notice the two sets have the same range

Standard Deviation And Variance For A Population

- *Variance* is the mean squared distance from the mean. It is denoted by σ^2 .
- *Standard deviation* is the square root of the variance .
 - * It is denoted by σ .
 - * It provides a measure of the standard distance from the mean.
 - * A small Standard deviation indicated that the scores are typically close to the mean, and a large standard deviation indicated that the scores are generally far from the mean.

The logical steps leading to the formulas for the variance and standard deviation are as follows:

- A deviation score is defined as $x - \mu$ and measures the direction and distance from the mean for each score.
- Because of the plus and minus sign, the sum of the deviation scores and the average of the deviation scores will always be zero.
- To get rid of the signs, we square each deviation and then compute the mean squared deviation, or the variance.
- Finally, we correct for having squared all the deviation by taking the square root of the variance. The result is the standard deviation, and it gives a measure of the standard distance from the mean.

Formulas

$$\begin{aligned}
 \text{variance} &= \text{mean squared error} \\
 &= \frac{\text{sum of squared deviation}}{\text{number of scores}} \\
 &= \frac{\sum (x - \mathbf{m})^2}{N} \\
 &= \frac{SS}{N}
 \end{aligned}$$

$$\text{population standard deviation} = \sqrt{\frac{SS}{N}}$$

Note that :

$$SS = \sum (x - \mathbf{m})^2 = \sum x^2 - \frac{(\sum x)^2}{N}$$

Standard Deviation And Variance For A Sample

The sample variance is denoted by s^2 .

The sample standard deviation is denoted by s .

$$\begin{aligned} \text{sample variance} &= \frac{\sum (x - \bar{x})^2}{n - 1} \\ &= \frac{SS}{n - 1} \end{aligned}$$

$$\text{sample standard deviation} = \sqrt{\frac{SS}{n - 1}}$$

Note that :

$$SS = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

Note that the sample formulas use $n-1$ instead of n . This is the adjustment necessary to correct for the bias in sample variability.

Example 13:

Find the sample standard deviation for the following 7 scores.

1, 6, 4, 3, 8, 7, 6

$$\sum x = 35$$

$$\sum x^2 = 211$$

$$SS = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= 211 - \frac{(35)^2}{7}$$

$$= 36$$

$$s^2 = \frac{SS}{n - 1} = \frac{36}{6} = 6$$

$$s = \sqrt{6} = 2.45$$

Exercises

12. For each of the following two population, you should be able to determine the standard deviation without doing any serious calculation.

a. Population I scores: 5, 5, 5, 5

b. Population II scores: 4, 6, 4, 6

13. Can SS ever have a value less than zero?

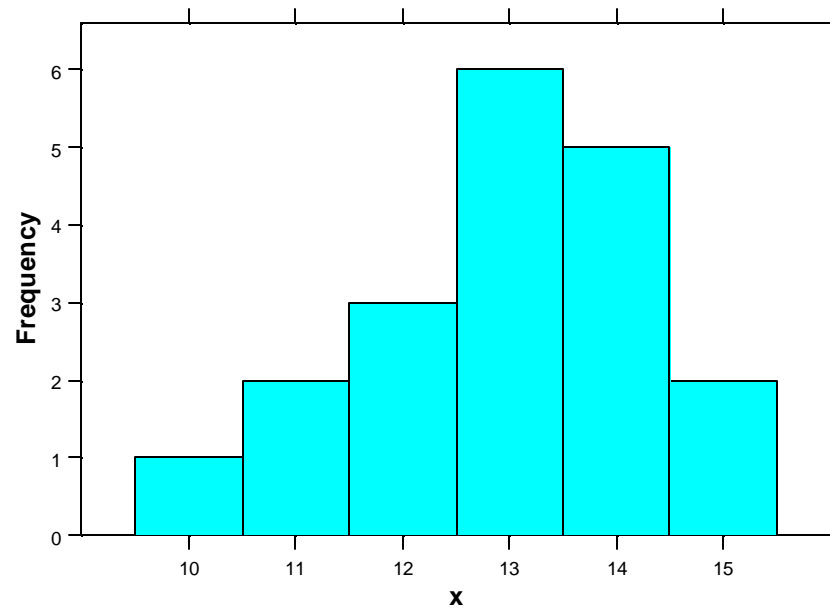
14. Calculate the standard deviation for the following Sample.

7, 9, 10, 8, 9, 12

15. A researcher is measuring student opinions using a standard 7-point scale (1= “strongly disagree” and 7=“strongly agree”). For one question, the researcher reports that the student responses averaged 5.8 with a standard deviation of 8.4. It should be obvious that the researcher made a mistake. Explain why.

Solution

1. It is left skewed.



2a.

0 0 7 9 9

1 0 3 4 6 8

2 3 6

3 0 1

4 2

5 2 8

6 1 2 2 6

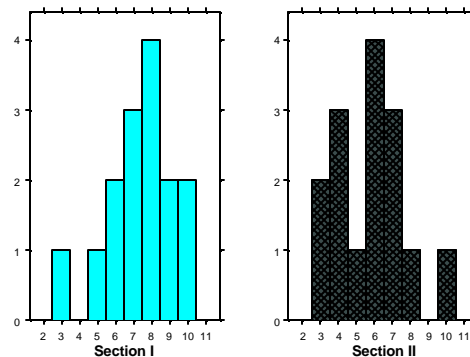
7 3 4 7 8 9

8 2

2.b 3

3. The bar graph is used when the data is measured in nominal or ordinal scale. While the histogram is used when the data is measured on an interval or ratio scale.

4a.



b. The scores in Section I are centered around $X=8$ and form a negatively skewed distribution. In Section II, the scores are centered around $X=6$ and there is a tendency toward a positively skewed distribution.

5. Since the extreme scores could displace the values of the mean. Also the mean may not represent the data.

6. In the case where we have a symmetric, bimodal distribution the mean and the median would not be representative of most of the scores. The individual scores would be clustered around the two modes with relatively few scores located in the center.

7. The mean, the median, and the mode all have the same value when the scores have a symmetric distribution with one mode.

8a.

$$\mathbf{m} = \frac{(6 + 2 + 4 + 5 + 3 + 7 + 6 + 2 + 1)}{9} = 4$$

To find the median,

Arrange the data 1, 2, 2, 3, 4, 5, 6, 6, 7 .

Find the location of the median=5.

So, 4 is the median.

8b.

$$\mathbf{m} = \frac{(6 + 2 + 4 + 5 + 3 + 7 + 6 + 2 + 1 + 100)}{10} = 13.6$$

To find the median,

Arrange the data 1, 2, 2, 3, 4, 5, 6, 6, 7, 100 .

Find the location of the median=5.5.

So, 4.5 is the median.

The median value is more resistant to the extreme values.

- 9.** When the score comes from a nominal scale.
- 10.** If the new value remains on the same side of the median (above or below) as the original value, then the median will not change.
- 11.** Negatively skewed.

12a. Since the scores are all the same, the variability is 0.

b. The mean of those scores is 5. The difference between each score and the mean is +1 or - 1 . So the square of each is 1 and the sum of the four squared differences is 4. Hence, the standard deviation is 1.

13. No.

14.

$$\sum x = 55$$

$$\sum x^2 = 519$$

$$SS = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= 519 - \frac{(55)^2}{6}$$

$$= 14.83$$

$$s^2 = \frac{SS}{n - 1} = \frac{14.83}{5} = 2.966$$

$$s = \sqrt{2.966} = 1.72$$

15. Assume that the average value the researcher got is true. Then the smallest distance from the scores to the mean is 0.2 and the largest is 4.8. Hence, the standard deviation should be between 0.2 and 4.8. As we can see 8.4 is not in that range.

Z-score

Location Of Scores
And
Standardized Distribution

Example:

In an exam a student received a score of 37, whereas the mean of the class is 28 with a standard deviation of 6. In another section, another student received a 46. The distribution for this section has a mean of 35 and a standard deviation of 10. Who has a higher standing in the class?

Solution:

Student I

$$\text{score} = 37$$

$$\text{mean} = 28$$

$$\text{standard deviation} = 6$$

The difference between the mean and the score $= 37 - 28 = 9$.
Now, 9 is **1.5** standard deviation above the mean.

Student II

score = 46

mean = 35

standard deviation = 10

The difference between the mean and the score = $46 - 35 = 11$. Now, 11 is **1.1** standard deviation above the mean.

Hence, Student I has a higher standard in the class.

The values 1.5 and 1.1 are called the z-score for 37 and 46 respectively.

Definition:

A z-score specifies the precise location of each X value within a distribution. The sign of the z-score (+ or -) signifies whether the score is above the mean (positive) or below the mean (negative). The numerical value of the z-score specifies the distance from the mean by counting the number of standard deviations between X and the mean.

Z - Score formula:

The relation between X and z-score can be expressed symbolically in a formula for transforming raw scores into z-scores is

$$z = \frac{x - m}{s}$$

w h e r e

m i s t h e m e a n o f t h e d i s t r i b u t i o n

s i s t h e s t a n d a r d d e v i a t i o n o f t h e
d i s t r i b u t i o n .

The numerator of the equation is the deviation score and measures the distance in points between X and μ .

We divide this difference by σ because we want the z-score to measure distance in terms of standard deviation units.

The Characteristic of a z-score

It is possible to describe the location of every raw score in the distribution by assigning z-scores to all of them. The result would be a transformation of the distribution of raw scores into a distribution of z-scores.

- The **shape** of the z-score distribution will be the same as the original distribution of the raw scores.
- The **mean** of the z-score will always be zero.
- The **standard deviation** of the z-score will always be 1.

Example:

For the following scores

9, 8, 7, 8, 9, 10, 7, 8, 9, 6, 8, 10

1. Find the z-score for each score
2. Draw a histogram for each of the raw scores and the z-scores to describe their distribution. Comment on the histogram.
3. Find the mean and the standard deviation for the z-scores.

Solution:

1. To find the z-scores we need to calculate the mean and the standard deviation for the y-scores

$$\text{mean}(y) = 8.25$$

$$\text{std}(y) = 1.215$$

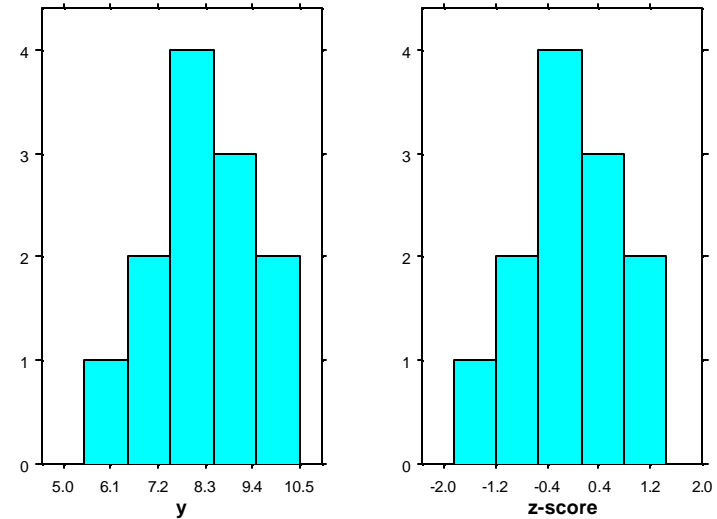
So, the z-core for $y=9$ is

$$\text{z-score} = (9-8.25)/1.215 = 0.62$$

and so on we get

<u>y</u>	<u>z-score</u>
9	0.62
8	-0.21
7	-1.03
8	-0.21
9	0.62
10	1.44
7	-1.03
8	-0.21
9	0.62
6	-1.85
8	-0.21
10	1.44

2. They look identical.



3. $\text{Mean}(\text{z-score}) = 0$

$\text{std}(\text{z-score}) = 1$

Exercises

1. The mean of a distribution after a z-score transformation is always zero because $\Sigma z = 0$. Explain why Σz must always equal zero.
2. For a distribution of raw scores, the mean $\mu = 45$. The z-score for $X = 55$ is computed and a value of $z = -2.00$ is obtained. Regardless of the value of the standard deviation, why must this z-score be incorrect.

NORMAL DISTRIBUTION

Characteristics Of the Normal Distribution

- It is a continuous distribution
- It is a symmetrical distribution
- It is asymptotic to the axis
- It is unimodal
- It is a family of curves
- Area under the curve is 1

- One of the most important properties of normal random variables is that within a fixed number of standard deviations from the mean, all normal distributions contain the same fraction of their probabilities

- Approximately 68% of the data is within 1 standard deviation of the mean
- Approximately 95% of the data is within 2 standard deviation of the mean
- Approximately 99% of the data is within 3 standard deviation of the mean

If **X** has a normal distribution with mean and standard deviation .
Then we write it

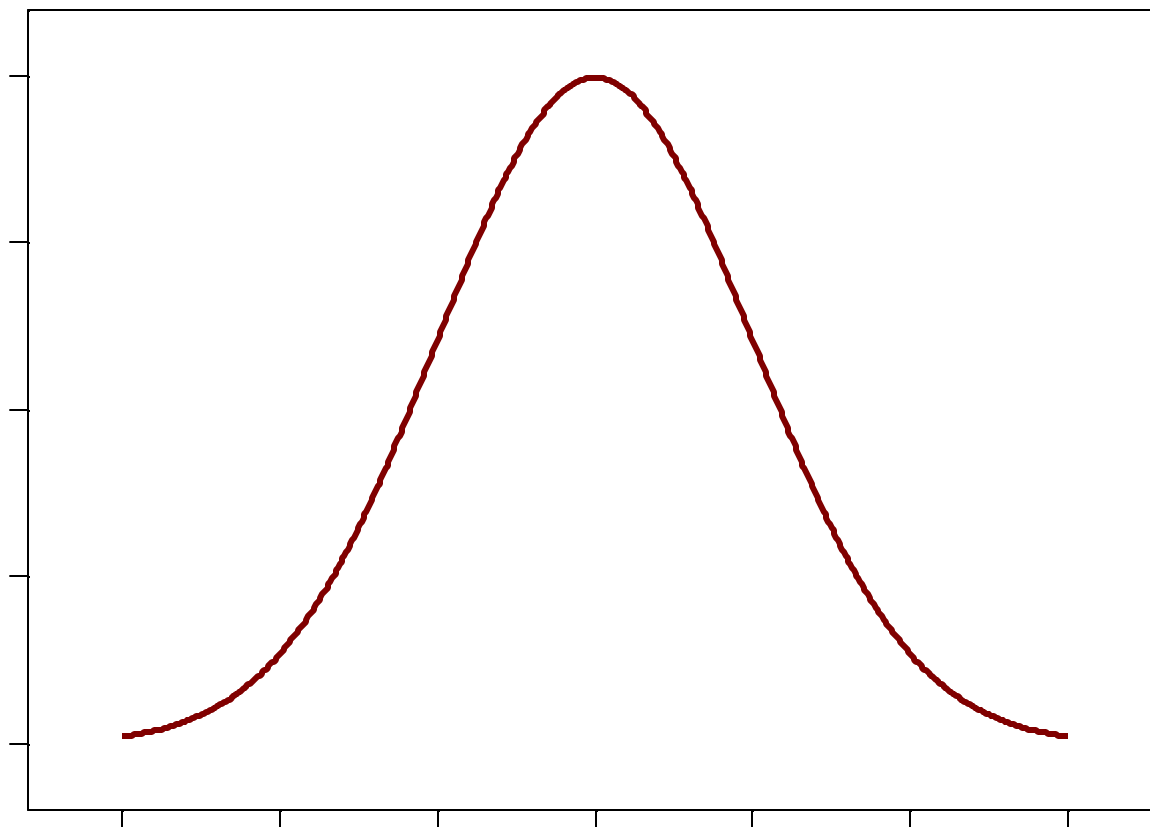
$$\mathbf{X} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}).$$

If $\boldsymbol{\mu} = 0$ and $\boldsymbol{\sigma} = 1$ the **X** is said to have a standard normal distribution and Usually is denoted by **Z**.

The mathematical equation that represent the distribution of **X** is:

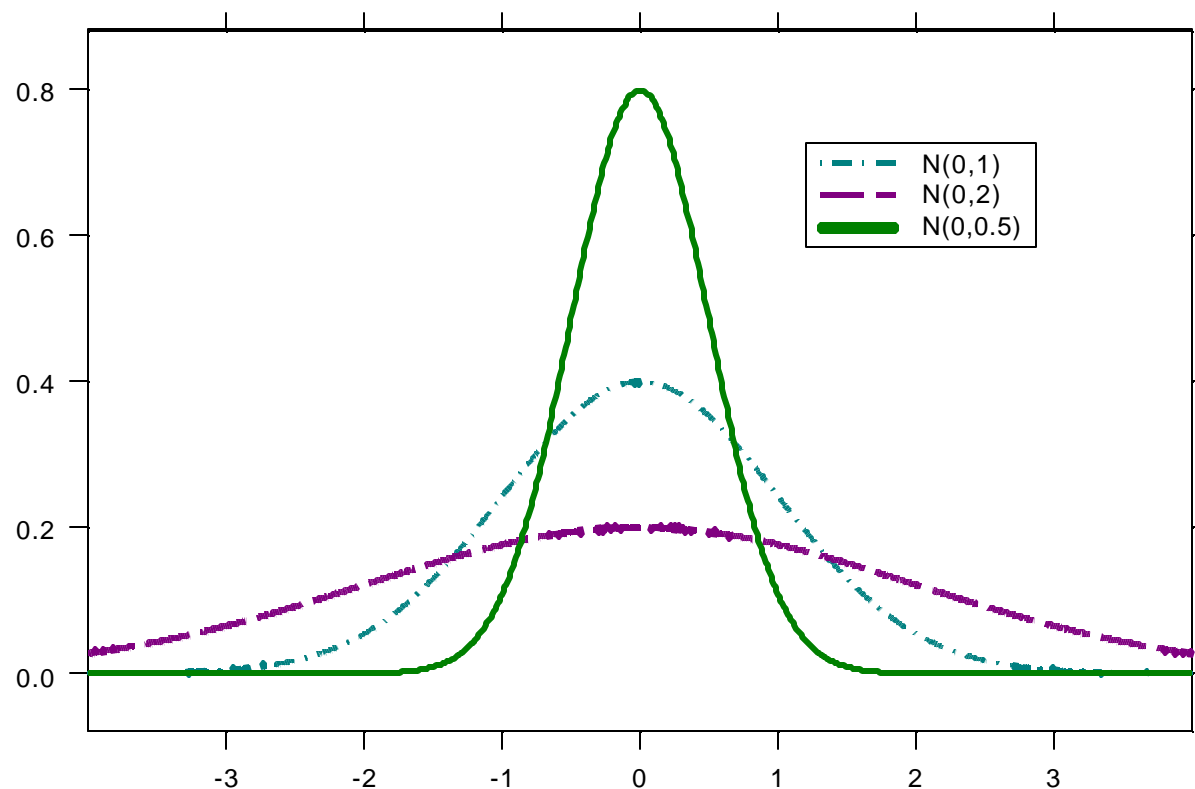
$$f (x / \mathbf{m}, \mathbf{s}) = \frac{1}{\sqrt{2\pi s^2}} e^{- (X - \mathbf{m})^2 / 2 s^2}$$

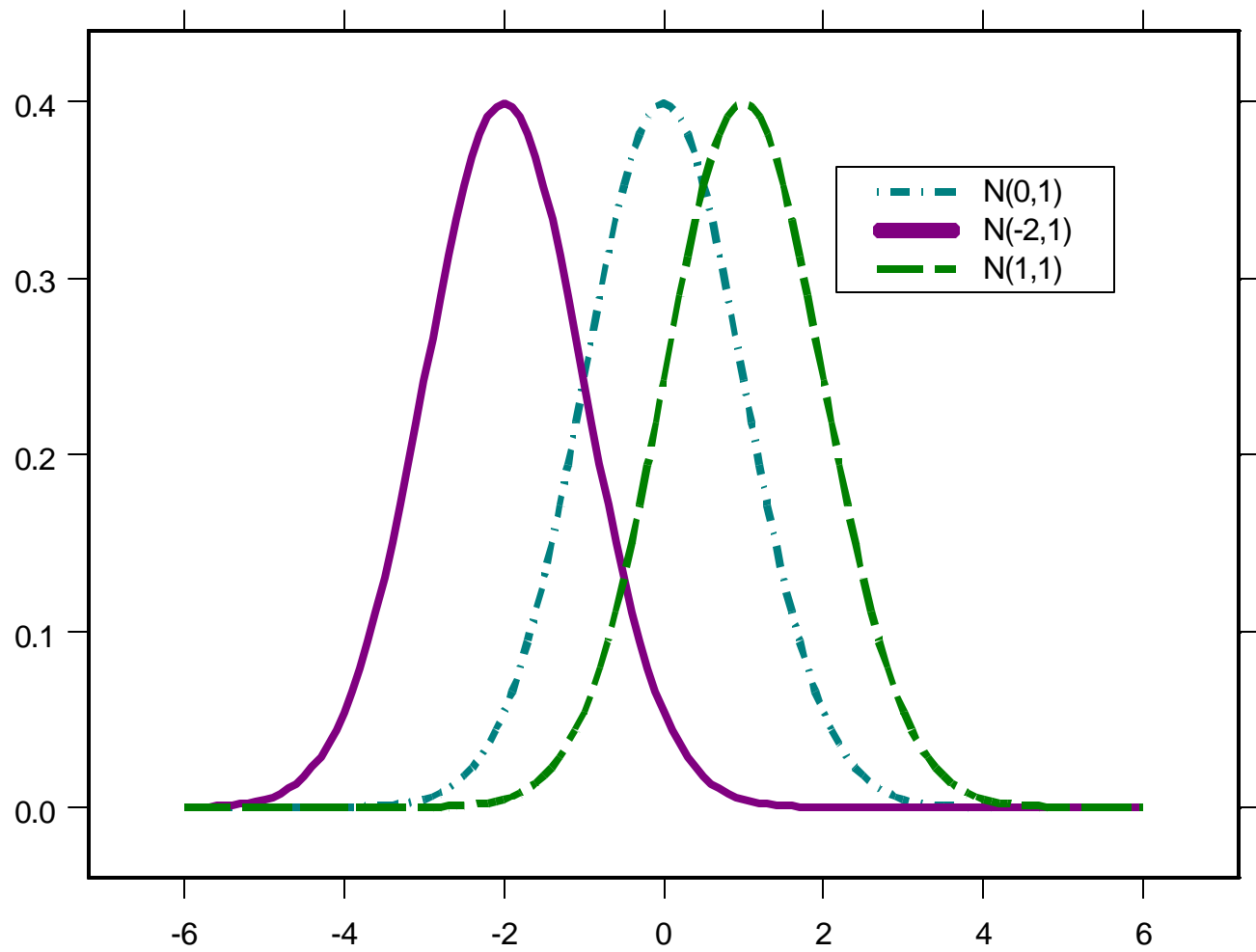
In general, the Normal distribution looks like this



Graphs of the Normal Distribution

Notice that the mean defines the location of the graph and the variance defines the dispersion.





Probability And Normal Distribution

- To use the probability density function to determine the probability of some interval would be complicated.
- The standard normal distribution provides a basis for computing probabilities for all normal distributions.
- The techniques used to translate any normal random variable into a standard normal random variable is using the z-score
- A table contains probability calculations for various points in the standard normal distribution is usually provided (see handout).
- The table provides the probability that a standard normal random variable will be between 0 and a specified value and the area beyond that value.

Example

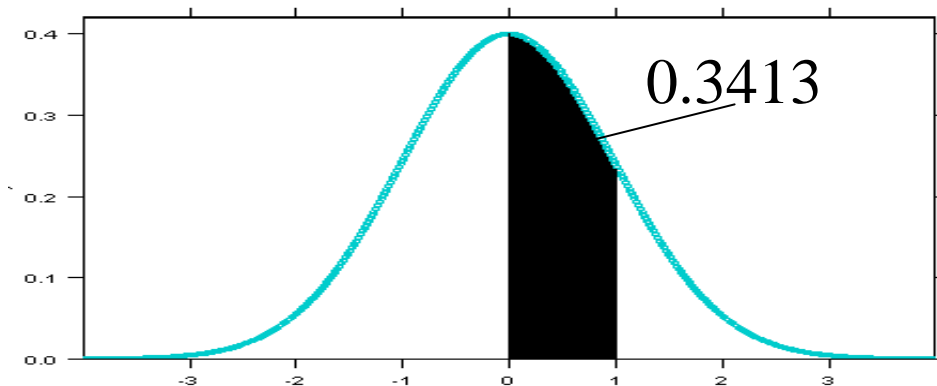
Find the probability that a standard normal random variable will be between 0 and 1?

To do so,

Draw the picture, then look up the value 1.00 in the table. The table value of 0.3413 is the area under the curve between 0 and 1, which is also the probability that the random variable will assume a value in that interval.

So ,

$$P(0 < Z < 1)$$



Example

Using the table, determine the probability that a standard normal random variable is between -1.32 and 0.

To do so,

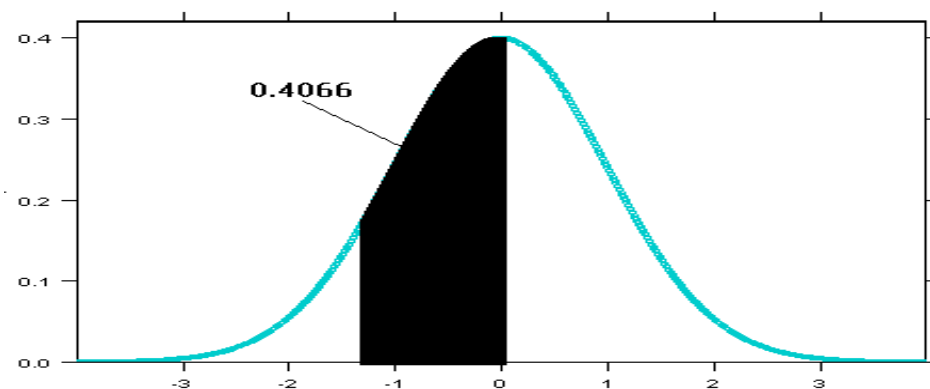
First, Draw the picture.

The value $z=-1.32$ is not found in the table. So we have to use the symmetric property of the Normal Distribution which implies that

$$P(-1.32 < z < 0) = P(0 < z < 1.32)$$

$$P(0 < z < 1.32) = 0.4066.$$

$$\text{Thus, } P(-1.32 < z < 0) = 0.4066.$$



Example

Using the table, determine the probability that a standard normal random variable is between 1 and 2?

To do so,

First, Draw the picture.

Determine the probability that z is between 0 and 2.0.

$$\mathbf{P(0 < z < 2.0) = .4772}$$

Determine the Probability that z is between 0 and 1.0.

$$\mathbf{P(0 < z < 1.0) = .3413}$$

Then,

$$\begin{aligned}\mathbf{P(1 < z < 2)} &= \mathbf{P(0 < z < 2) - P(0 < z < 1)} \\ &= \mathbf{.4772 - .3413} \\ &= \mathbf{.1359}\end{aligned}$$

Example

Adult height form a normal shaped distribution with a mean of 68 inches and a standard deviation of 6 inches. Given this information about the population, find the probability of randomly selecting an individual whose height is between 68 and 74 inches?

Let X be the adult height. Then $X \sim N(68,6)$

So we need to find

$$P(68 < X < 74) = \text{??????}$$

To get the answer we need to transform the distribution of X into the Standard Normal distribution. I.e. we need to transform it to a Random Variable that has a mean of 0 and a standard deviation of 1. To do so, we use the z-score transformation.

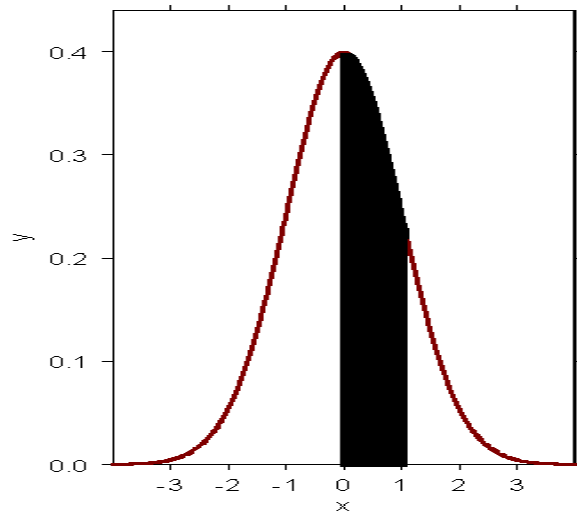
Find the z-score for 68 and 74 as follows:

For $x = 68$, the z-score $= (68-68)/6 = 0$.

For $x = 74$, the z-score $= (74-68)/6 = 1$.

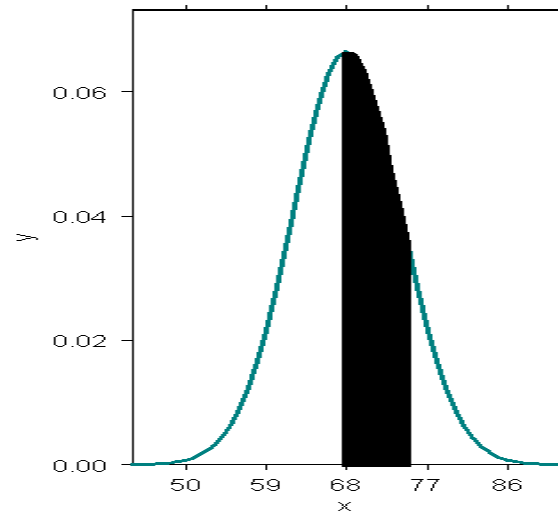
Hence,

$$P(68 < X < 74) = P(0 < Z < 1) = 0.3413$$



$$\mu = 0$$

$$\mu + 1.0 * \sigma = 1.0$$



$$\mu = 68$$

$$\mu + 1.0 * \sigma = 74$$

Exercises

3. For a normal distribution with a mean of 80 and a standard deviation of 10, find each probability value requested.

a. $P(X > 85)$

b. $P(X > 70)$

c. $P(X < 95)$

d. $P(75 < X < 100)$

4. For a normal distribution with a mean of 100 and a standard deviation of 20, find each value requested.

a. What score separate the top 40% from the bottom 60% of the distribution?

- b. What is the minimum score needed to be in the top 5% of this distribution?
- c. What scores from the boundaries for the middle 60% of this distribution?

5. A Cola distributor believes that the amount of Cola in a 12 ounce can of cola has a normal distribution with a mean of 12 ounces and a standard deviation of 1 ounce. If a 12 ounce Cola can is randomly selected, find the following probabilities.

- a. Find the probability that the 12 ounce can of Cola will actually contain less than 11 ounces of Cola.
- b. Find the probability that the 12 ounce can of Cola will actually contain more than 12.5 ounces of Cola.
- c. Find the probability that the 12 ounce can of Cola will actually contain between 10.5 and 11.5 ounces of Cola.

6. The Trial Making Test is frequently used by clinical psychologists to test for brain damage. Patients are required to connect consecutively numbered circles on a sheet of paper. It has been determined that the mean length of time required for a patient to perform this task is 32 seconds and the standard deviation is 4 seconds. Assume the distribution of the length of time required to connect circles is normal

a. Find the probability that a randomly selected patient will take longer than 40 seconds to perform the task.

b. Find the probability that a randomly selected patient will take between 24 and 40 seconds to complete the task.

c. A psychologist would like to retest those persons with completion times in the highest 5% of the distribution of times required. What time would a person need to exceed on the Trial Making Test to be considered for retesting?

The Sample Mean

Example

Consider a population that consists of only four scores: 2, 4, 6, 8. Construct the sampling distribution of the sample mean for $n = 2$?

Note that:

The mean of the population $\mu = 5$.

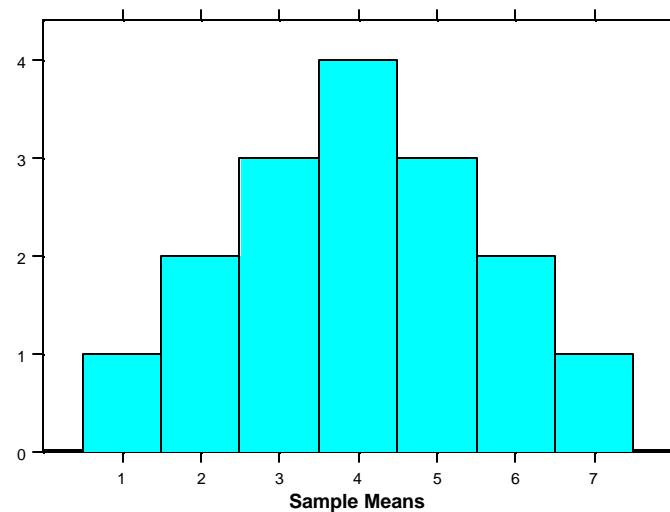
The standard deviation of the population $\sigma_{\bar{x}} = \sqrt{5}$

Now, Let us construct the Sampling Distribution of the Sample Mean for $n= 2$. To do so we need to find the Sample Mean for all possible random samples of size $n=2$.

Sample	Scores		Sample Mean
	First	Second	
1	2	2	2
2	2	4	3
3	2	6	4
4	2	8	5
5	4	2	3
6	4	4	4
7	4	6	5
8	4	8	6
9	6	2	4
10	6	4	5
11	6	6	6
12	6	8	7
13	8	2	5
14	8	4	6
15	8	6	7
16	8	8	8

We notice the following:

- The sample means are clustered around the value 5 which represents the population mean.
- The distribution of the sample means is approximately normal in shape.
- The mean of the 16 sample means $\bar{X} = 5$
- The standard deviation of the sample means $s_{\bar{x}} = \frac{\sqrt{5}}{\sqrt{2}}$



Definition:

The distribution of the sample mean is the collection of sample means for all possible random samples of a particular size that can be obtained from the population.

Definition:

A sampling distribution is a distribution of a statistics obtained by selecting all possible samples of a specific size of a population.

Definition: Central Limit Theorem

For any population with mean μ , and standard deviation σ , the distribution of sample means for a sample size n will approach a normal distribution with a mean μ and a standard deviation of $s_x = \frac{s}{\sqrt{n}}$ as n approaches infinity .

The central limit Theorem describes the distribution of sample means for any population, no matter what shape, or mean, or standard deviation. Also, the distribution of the sample means “approaches” a normal distribution very rapidly ($n=30$)

The Shape of the distribution of the sample mean

By the central limit theorem, the distribution tends to be a normal distribution. In fact it will be almost perfectly normal if either one of the following is satisfied

1. The population from which the samples are selected is a normal distribution.
2. The number of scores (n) in each sample is relatively large, around 30 or more.

The mean of the distribution of Sample Means

The mean of the distribution of the sample mean will be equal to μ and is called the expected value of the sample mean.

The standard error of the distribution of Sample Means

The standard deviation of the distribution of the sample mean is called the standard error.

Example

On an immediate memory test, 10-year-old children can correctly recall an average of $\mu = 7$ digits. The distribution of recall scores is normal with $\sigma = 2$.

a. What is the probability of randomly selection a child with a recall score less than 6?

b. What is the probability of randomly selecting a sample of $n=4$ children whose average recall score is less than 6.

C. Same as in b but with a sample size of 9

Solution

a. Let X : Recall Scores then $X \sim N(7,2)$

$$\begin{aligned} P(X < 6) &= P(Z < (6-7)/2) \\ &= P(Z < -0.5) \\ &= P(Z \geq 0.5) \\ &= 0.3085 \end{aligned}$$

b. Let \bar{X} : Average Recall Scores for $n=4$ then $\bar{X} \sim N(7,1)$

$$\begin{aligned} P(\bar{X} < 6) &= P(Z < (6-7)/1) \\ &= P(Z < -1) \\ &= P(Z \geq 1) \\ &= 0.1587 \end{aligned}$$

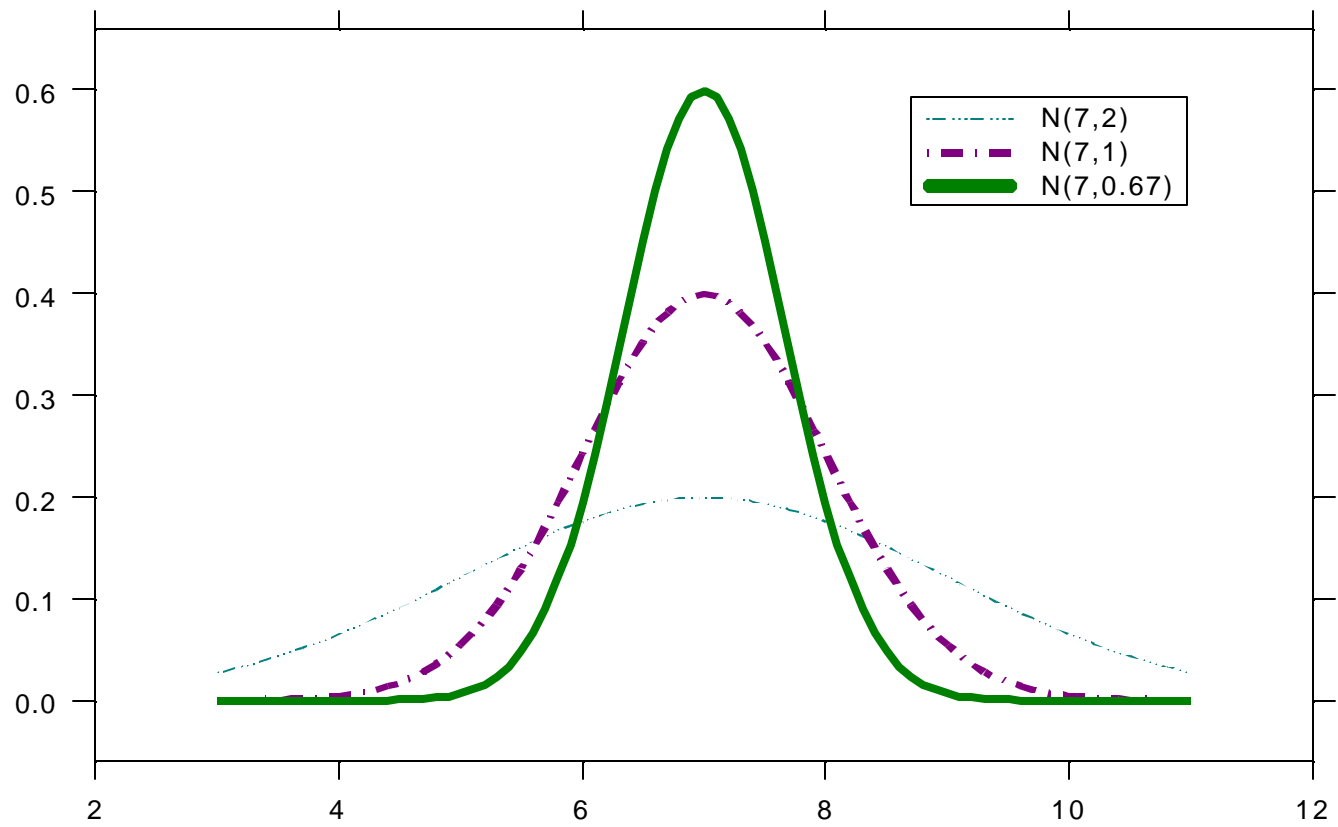
c. Let \bar{x} : Average Recall Scores for $n=9$ then $\bar{x} \sim N(7, 0.67)$

$$P(\bar{x} < 6) = P(Z < (6-7)/0.67)$$

$$= P(Z < -1.5)$$

$$= P(Z \geq 1.5)$$

$$= 0.0668$$



Exercises

7. A large grocery chain in New York has received a shipment of 1000 cases of oranges. The shipper claims that the cases average $\mu = 40$ oranges with a standard deviation of $\sigma = 2$. To check this claim, the store manager randomly selects 4 cases and counts the number of oranges in each case. For these 4 cases, the average number of oranges is $\bar{X} = 38$

- a. Assuming that the shipper's claim is true, what is the probability of obtaining a sample mean this small.
- b. Based on your answer for part a, does the grocery store manager have a reason to suspect that he has been cheated? Explain your answer.

8. IQ scores from a normal distribution with $\mu = 100$ and $\sigma = 15$

a. What is the probability of randomly selecting a sample of $n=9$ people so that their average IQ is different by more than one point from the population mean?

b. What is the probability of randomly selecting a sample of $n=100$ individuals so that their average IQ is more than one point away from the population mean?

Solution

1. We need to show that $\sum z = 0$.

$$\begin{aligned}
 \text{Since } z &= \frac{x - m}{s} \\
 \sum_{i=1}^n z &= \sum_{i=1}^n \left(\frac{x - m}{s} \right) \\
 &= \frac{1}{s} \sum_{i=1}^n (x - m) \\
 &= \frac{1}{s} \left(\sum_{i=1}^n x - \sum_{i=1}^n m \right) \\
 &= \frac{1}{s} \left(\sum_{i=1}^n x - (n m) \right) \\
 &= \frac{1}{s} (n m - n m) \\
 &= 0
 \end{aligned}$$

2. The value of the z-score should be positive since the score $x = 55 > \mu = 45$.

3. So, $X \sim N(80, 10)$

a.
$$\begin{aligned} P(X > 85) &= P\left(\frac{X - 80}{10} > \frac{85 - 80}{10}\right) \\ &= P(Z > 0.5) \\ &= 0.3085 \end{aligned}$$

b.
$$\begin{aligned} P(X > 70) &= P\left(\frac{X - 80}{10} > \frac{70 - 80}{10}\right) \\ &= P(Z > -1) \\ &= P(Z \leq 1) \\ &= 0.5 + 0.3413 \\ &= 0.8413 \end{aligned}$$

c.

$$\begin{aligned}P(X < 95) &= P\left(\frac{X - 80}{10} < \frac{95 - 80}{10}\right) \\&= P(Z < 1.5) \\&= 0.5 + 0.4332 \\&= 0.9332\end{aligned}$$

d.

$$\begin{aligned}P(75 < X < 100) &= P\left(\frac{75 - 80}{10} < \frac{X - 80}{10} < \frac{95 - 80}{10}\right) \\&= P(-0.5 < Z < 2) \\&= P(-0.5 < Z < 0) + P(0 < Z < 2) \\&= P(0 < Z < 0.5) + P(0 < Z < 2) \\&= 0.1915 + 0.4772 \\&= 0.6687\end{aligned}$$

4. So, $X \sim N(100, 20)$. Now to find the value of the x 's we need to find the z -score corresponds for each first and then use the equation $x = \sigma z + \mu$.

a. $P(Z > z) = 0.4$ so,

$$\text{So, } z = 0.255$$

$$\text{Hence, } x = 20(0.255) + 100 = 105.1$$

b. $P(Z > z) = 0.05$

$$\text{So, } z = 1.645$$

$$\text{Hence, } x = 20(1.645) + 100 = 132.9$$

c. $P(-z < Z < z) = 0.6$

$$\text{So, } P(0 < Z < z) = 0.3$$

$$\text{So, } z = 0.845$$

$$\text{Hence, } x = 20(0.845) + 100 = 116.9$$

5. Let,

X : the amount of Cola in a 12 ounce can

$X \sim N(12,1)$

a.

$$\begin{aligned} P(X < 11) &= P\left(\frac{X - 12}{1} < \frac{11 - 12}{1}\right) \\ &= P(Z < -1) \\ &= P(Z > 1) \\ &= 0.1587 \end{aligned}$$

b.

$$\begin{aligned} P(X > 12.5) &= P\left(\frac{X - 12}{1} > \frac{12.5 - 12}{1}\right) \\ &= P(Z > 0.5) \\ &= 0.3085 \end{aligned}$$

c.

$$\begin{aligned}P(10.5 < X < 11.5) &= P\left(\frac{10.5 - 12}{1} < \frac{X - 12}{1} < \frac{11.5 - 12}{1}\right) \\&= P(-1.5 < Z < -0.5) \\&= P(0.5 < Z < 1.5) \\&= P(0 < Z < 1.5) - P(0 < Z < 0.5) \\&= 0.4332 - 0.1915 \\&= 0.2417\end{aligned}$$

6. Let, X: Time required for the patient to connect consecutively numbered circles

$$X \sim N(32, 4)$$

$$\begin{aligned}\text{a. } P(X > 40) &= P\left(\frac{X - 32}{4} > \frac{40 - 32}{4}\right) \\&= P(Z > 2) \\&= 0.0228\end{aligned}$$

b.

$$\begin{aligned}P(24 < X < 40) &= P\left(\frac{24 - 32}{4} < \frac{X - 32}{4} < \frac{40 - 32}{4}\right) \\&= P(-2 < Z < 2) \\&= 2P(0 < Z < 2) \\&= 2 * 0.4772 \\&= 0.9544\end{aligned}$$

c. We need to find x such that $P(X > x) = 0.05$. To do so we find z such that $P(Z > z) = 0.05$ and then find x .

From the table $z = 1.645$ so

$$x = (4 * 1.645) + 32 = 38.58.$$

7. Let X = number of oranges in each case.

We know that $\mu = 40$ oranges and the standard deviation of $\sigma = 2$ and $n = 4$

So $\bar{X} \sim N(40, \frac{2}{\sqrt{4}}) = N(40, 1)$

a.

$$\begin{aligned} P(X < 38) &= P\left(\frac{X - 40}{1} < \frac{38 - 40}{1}\right) \\ &= P(Z < -2) \\ &= P(Z > 2) \\ &= 0.0228 \end{aligned}$$

b. The manager should suspect that he has been cheated. It is extremely unlikely to obtain this sample if the shipper's claim is true.

8. Let X = IQ scores. $X \sim N(100, 15)$

$$n = 9$$

$$\text{So, } \bar{X} \sim N(100, \frac{15}{\sqrt{9}}) = N(100, 5)$$

a.

$$\begin{aligned} P(X < 99) + P(X > 101) &= \\ P\left(X < \frac{99 - 100}{5}\right) + P\left(X > \frac{101 - 100}{5}\right) &= \\ = P(Z < -0.20) + P(Z > 0.20) &= \\ = 2 * P(Z > 0.20) &= \\ = 2 * 0.4207 &= \\ = 0.8414 \end{aligned}$$

b.

$$\overline{X} \sim N\left(100, \frac{15}{\sqrt{100}}\right) = N(100, 1.5)$$

$$\begin{aligned} P(X < 99) + P(X > 101) &= \\ P\left(X < \frac{99 - 100}{1.5}\right) + P\left(X > \frac{101 - 100}{1.5}\right) &= \\ = P(Z < -0.67) + P(Z > 0.67) &= \\ = 2 * P(Z > 0.67) &= \\ = 2 * 0.2514 &= \\ = 0.5028 \end{aligned}$$

Statistical Inference

Hypothesis Testing For Single Population

Objectives:

- Understand the logic of hypothesis testing
- How to establish the null and the alternative hypothesis
- Understand Type I and Type II errors.
- Statistical power.
- Test the hypothesis about the population mean when σ is known.
- Test the hypothesis about the population mean when σ is unknown assuming that the population is normally distributed.

The procedure of *hypothesis testing* is useful in making decisions about a parameter value. For example, we may be interested in deciding whether the mean tar content, μ , of a particular brand of cigarette exceeds a certain value may be 4 milligrams; whether the mean lifetime of a product manufactured by industry A is less than the mean lifetime of a similar product manufactured by industry B; whether the proportion of American who believe that the president is doing a good job exceeds 0.5; etc..

Steps in Testing Hypotheses

- Establish hypothesis: state the null and the alternative Hypothesis.
- Determine the appropriate statistical test and sampling distribution.
- Specify Type I error rate.
- State the decision rule.
- Gather Sample data.
- Calculate the value of the test statistic.
- State the statistical conclusion.

Null and Alternative Hypotheses

A statistical hypothesis is a statement about the value of a population parameter. To establish this statement we need to establish two hypothesis that are set up in opposition to each other. They are:

Null Hypothesis: Denoted by H_0 and it is the hypothesis which we hope to gather information against.

Alternative Hypothesis: Denoted by H_1 and it is the hypothesis we wish to gather supporting evidence.

Example 1:

A medical researcher would like to determine whether the proportion of males admitted to a hospital because of heart disease differs from the corresponding proportion of females.

The Hypothesis must be stated in terms of a population parameter or parameters. Lets,

ρ_1 = the proportion of men admitted to a hospital because of heart disease.

ρ_2 = the proportion of women admitted to a hospital because of heart disease.

The researcher wants to support the claim that ρ_1 is different than ρ_2 , therefore the null and the alternative hypothesis, in terms of those parameters are:

$$H_0 : \rho_1 = \rho_2$$

$$H_1 : \rho_1 \neq \rho_2$$

Example 2:

Clinical researchers speculate that children who drink milk fortified with calcium tend to develop stronger and denser bones as adults and, consequently, are less likely to suffer from osteoporosis. One study, conducted was designed to test whether the mean density of bones of women who drank milk with each meal as children is greater than the mean density of women who drank milk less frequently.

Let,

μ_1 = the mean density of women who drank milk with each meal as children.

μ_2 = the mean density of women who did not drink milk with each meal as children.

The null and the alternative hypothesis are given as follows:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

One-Tailed and Two-Tailed Tests

A **one-tailed test** of hypothesis is one in which the alternative hypothesis is directional as in Example 2.

A **two-tailed test** of hypothesis is one in which the alternative does not specify departure from the null hypothesis in a particular direction as in Example 1.

Acceptance and Rejection Region

After establishing the null and the alternative hypothesis , the researcher can set up decision rules to determine whether the null hypothesis is going to be rejected or not.

Example 3:

A metal Lathe is checked periodically by quality control inspectors to determine if it is producing machine bearings with a mean diameter of 0.5 inch. If the mean diameter of the bearings is larger or smaller than 0.5 inch, then the process is out of control and needs to be adjusted.

Let,

μ = True mean diameter in inches of all bearings produced by lathe

If either $\mu > 0.5$ or $\mu < 0.5$, then the metal lath's production process is out of control. So the null and the alternative hypothesis is as follows:

$H_0 : \mu = 0.5$ (the process is in control)

$H_1 : \mu \neq 0.5$ (the process is out of control)

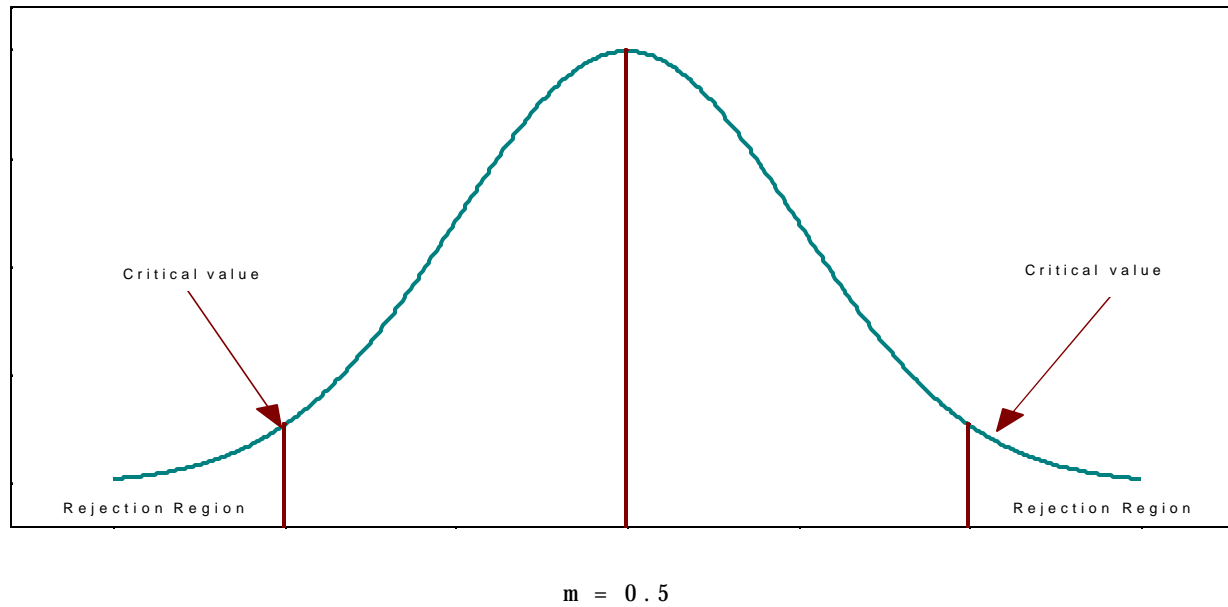
Common sense says that expecting all bearings produced to have a mean value of exactly 0.5 is unrealistic . In this way the hypotheses are structured around the mean values, not individuals.

Would the null hypothesis be rejected if a sample mean of 0.49 is obtained?

In testing a hypothesis, the researcher should establish a **rejection region** after determining the null and the alternative hypothesis.

Suppose that the sample mean of the bearings has a normal distribution. The rejection regions is established in the tails of the distribution because the only way to reject the null $H_0 : \mu = 0.5$ is to get results in the region of $\mu \neq 0.5$.

Each rejection region is divided from the rest of the distribution by a point called **critical value**.



If the results obtained from the sample data yield a computed value in the rejection region, the null hypothesis is rejected.

Type I and Type II Errors

- **Type I error** occurs if we reject a null hypothesis when it is true.

- The probability of committing a Type I error is denoted by **α** and is called the **level of significance** (or the significance level) for a hypothesis test.

- **Type II error** occurs if we fail to reject the null hypothesis when it is false.

- The probability of Type II error is denoted by **β** .

State of nature

Null true

Null false

Action

Fail to
Reject
null

Correct
decision

Type II
Error
(β)

Reject
Null

Type I
error
(α)

Correct
decision

In example 3, a **Type I error** occurs if we conclude that the process is out of control when in fact the process is in control, I.e., if we conclude that the mean bearing diameter is different from 0.5 inch, when in fact the mean is equal to 0.5 inch. The consequences of making such error would be that unnecessary time and effort would be expended to repair the metal lathe.

A **Type II error** occurs if we conclude that the mean bearing diameter is equal to 0.5 inch when in fact the mean differs from 0.5 inch. The consequence of making a Type II error is the metal Lathe would not be repaired when in fact the process is out of control.

Example 4:

The logic used in hypothesis testing has often likened to that used in the courtroom in which a defendant is on a trial for committing a crime.

- a.** Formulate appropriate null and alternative hypotheses for judging the guilt or innocence of the defendant.
- b.** Interpret Type I and Type II error in this context.
- c.** If you were the defendant, would you want α to be small or large? Explain.

Solution

a. Under our judicial system, a defendant is “innocent until proven guilty.” That is the burden of proof is not on the defendant to prove innocence; rather, the court must collect sufficient evidence to support the claim that the defendant is guilty. Thus, the null and alternative hypotheses would be

H_0 : Defendant is innocent

H_1 : Defendant is guilty

b. Type I error would be to conclude that the defendant is guilty, when the defendant is innocent.

Type II error would be to conclude that the defendant is innocent, when in fact the defendant is guilty.

		Actual Situation	
		Did not Commit Crime	Committed Crime
Jury's Verdict	Innocent	Verdict Correct	Type II Error (β)
	Guilty	Type I error (α)	Verdict Correct

c. I would definitely want α to be as small as possible. Since the consequences are so serious.

Statistical Power

Definition

The power of a statistical test is the probability that the test will **correctly reject a false null hypothesis.**

The power is the probability of obtaining sample data in the critical region when the null hypothesis is false.

$$\text{Power} = P(\text{reject a false } H_0) = 1 - b$$

Testing Hypothesis About A Single Mean

CASE I: σ is known

Example 5:

IQ scores for the general population from a normal distribution with $\mu = 100$ and $\sigma = 15$. However, there are data that indicate that children's intelligence can be affected if their mothers have German measles during pregnancy. Using hospital records, a researcher obtained a sample of $n = 20$ school children whose mothers all had German measles during their pregnancies. The average IQ for this sample was $\bar{x} = 97.3$. Do these data indicate that German measles have a significant effect on IQ? Test with $\alpha = 0.05$.

Step 1 *State the Hypothesis*

$H_0 : \mu = 100$ (there is no affect)

$H_1 : \mu \neq 100$ (there is affect)

Step 2 *Locate the rejection region*

To find the rejection region we need to find the distribution of the sample mean. Since the population is normal $\mu = 100$ and $\sigma = 15$ then,

$$\bar{x} \approx N(100, \frac{15}{\sqrt{20}})$$

with $\alpha = 0.05$, we want to identify the most unlikely 5% of this distribution. The most unlikely part of the normal distribution is on the tails. Therefore we divide our alpha level evenly between the tails, 2.5% or $p = 0.025$ per tail.

The z score corresponding to $p=0.025$ is $z = 1.96$. So the boundaries of the critical region are -1.96 on the left side and 1.96 on the right side.

Step 3 *Obtain the sample data and compute the test statistic*

The researcher obtained a sample mean of $\bar{x} = 97.3$

so, the corresponding z-score is -0.81

Step 4 *Make a decision about the null hypothesis*

The z-score that was obtained is not in the critical region hence, the sample mean the researcher found is not an unusual value. So we fail to reject the null hypothesis.

Testing Hypothesis About A Single Mean

CASE I: s is unknown

When σ is unknown, the standard error cannot be computed, and a hypothesis test based on the z-score is impossible.

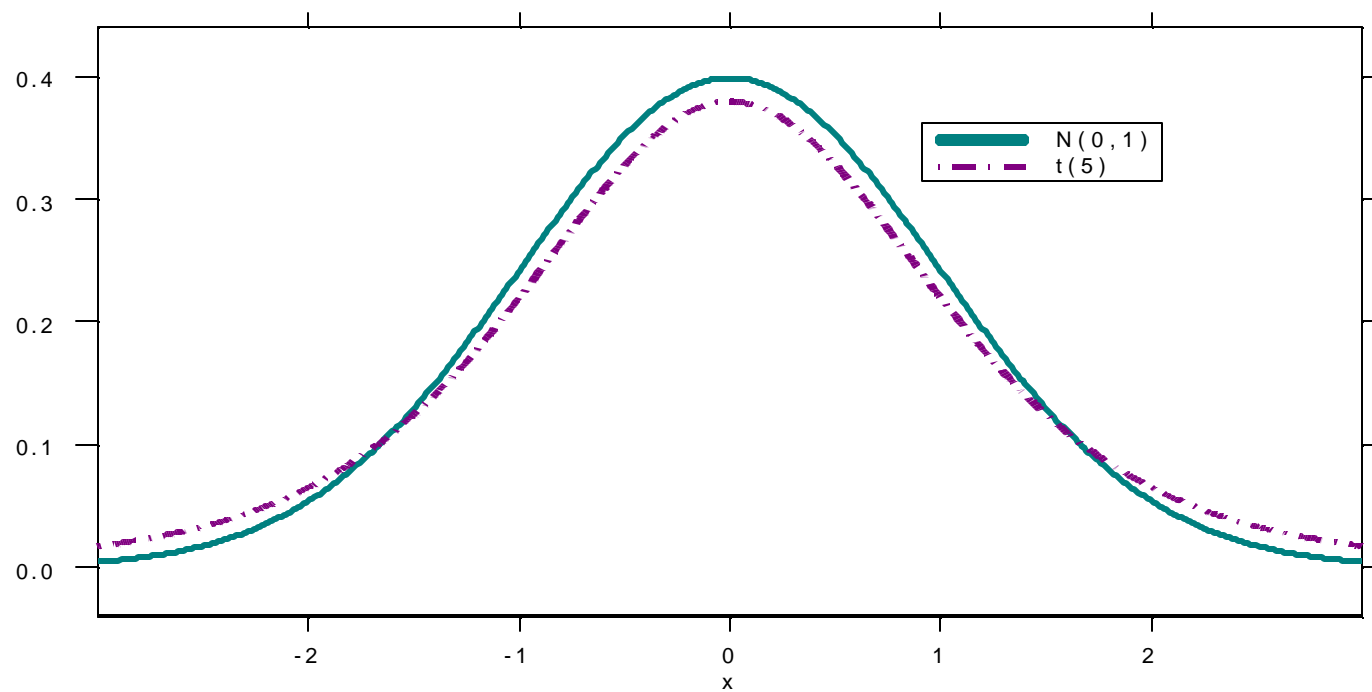
To test a hypothesis about μ when σ is unknown, σ must first be estimated using the sample standard deviation s . Next, the standard error is estimated by substituting s for σ in the standard error formula. Now, substitute the estimated standard error in the denominator of the z-score formula. The resulting test statistic is called a *t statistic*.

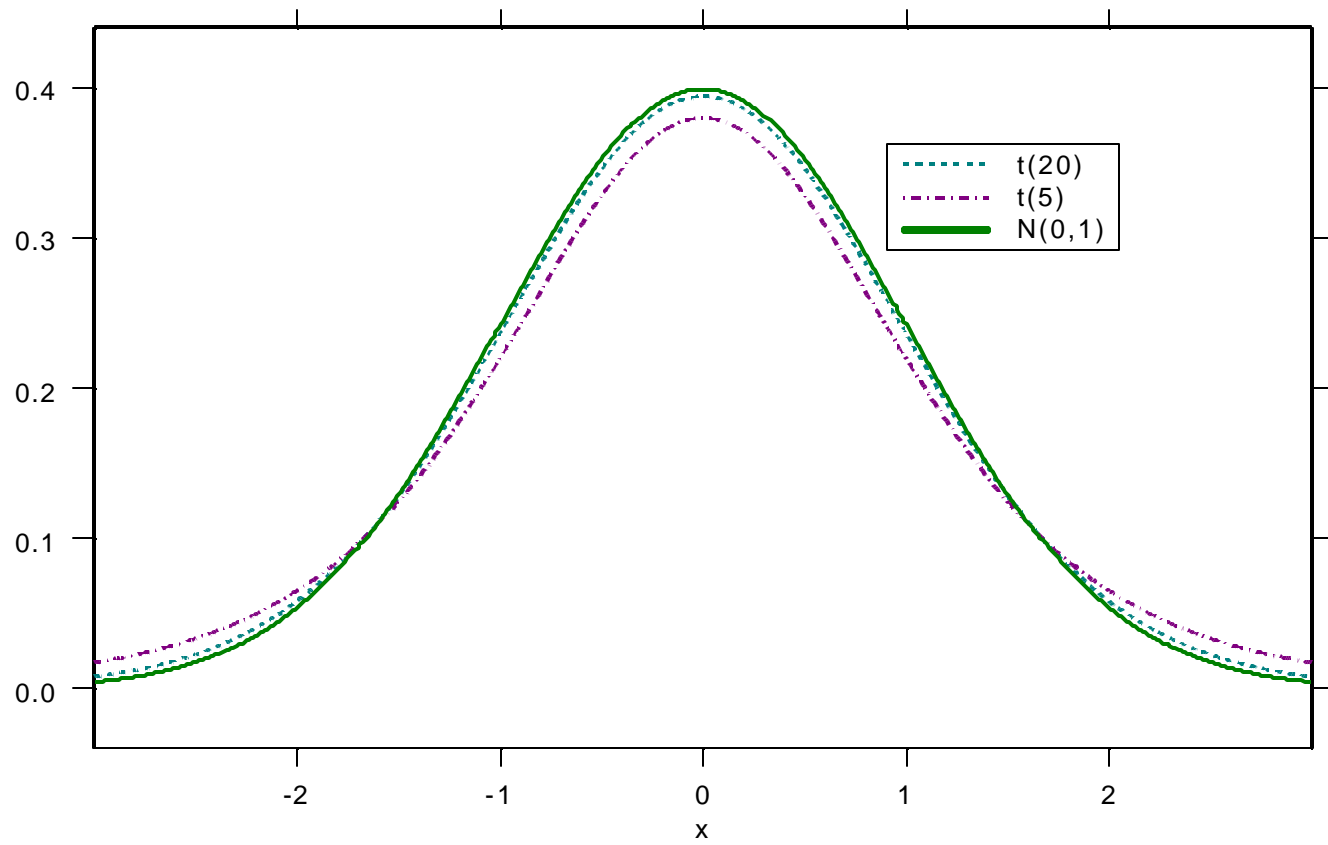
This ***t statistic*** has a t-distribution with n-1 degrees of freedom.

$$z = \frac{\bar{X} - m}{s_{\bar{X}}} = \frac{\bar{X} - m}{s / \sqrt{n}}$$

$$t = \frac{\bar{X} - m}{s_{\bar{X}}} = \frac{\bar{X} - m}{s / \sqrt{n}}$$

T- Distribution





To find the probabilities for the t-distribution we need to use a table. The two rows at the top of the table shows proportion of the t-distribution contained in either one or two tails, depending on which row is used. The first column is the degrees of freedom of the t statistic. The numbers in the body of the table are the t-values that mark the boundary between the tails and the rest of the t distribution

Example 6:

For $df=10$, what t values (s) are associated with

- a.** Top 1% of the t -distribution
- b.** the bottom 5% of the t distribution?
- c.** The most extreme 1% of the distribution

Solution

- a.** +2.764
- b.** -1.812
- c.** +3.169 and -3.169

Example 7:

A professor hypothesizes that an introductory course in logic will help college students with their other studies. To test this hypothesis, a random sample of $n=25$ freshman is selected. These students are required to complete a logic course during their freshman year. At the time of graduation, the final point average is computed for each of these students. The mean GPA for this sample is $\bar{x} = 2.83$ with $SS = 6$. Can the professor conclude that the grades for the sample were significantly different from the rest of the graduating class, which had an average GPA of $\mu = 2.58$? Test this with two-tailed test at $\alpha = 0.05$.

Step 1 *State the Hypothesis*

$$H_0 : \mu = 2.58$$

$$H_1 : \mu \neq 2.58$$

Step 2 *Locate the rejection region*

To find the rejection region we need to find the degrees of freedom . Since the sample size is 25 the $df=25-1 = 24$. Since, $\alpha =0.05$ and we have a two tailed test then the critical region begins at the t-values of +2.064 and - 2.064.

Step 3 *Obtain the sample data and compute the test statistic*

The professor obtained a sample mean of $\bar{x} = 2.83$

so, the corresponding t-score

$$\begin{aligned}
 t &= \frac{\bar{X} - m}{s_{\bar{X}}} \\
 &= \frac{\bar{X} - m}{s / \sqrt{n}} \\
 &= \frac{2.83 - 2.58}{0.5 / \sqrt{25}} \\
 &= 2.5
 \end{aligned}$$

Step 4 *Make a decision about the null hypothesis*

The t-score that was obtained is in the critical region hence, the sample mean the professor found is an unusual value. So we reject the null hypothesis.

Exercises

- 1.** After several years of studying human performance in flight simulators, a psychologist knows that the reaction times to an overhead emergency indicator from a normal distribution with $\mu = 200$ milliseconds and $\sigma = 20$. The psychologist would like to determine if placing the indicator in front of the person at eye level has any effect on reaction time. A random sample of $n = 25$ people is selected, they are tested in a simulator with the indicator light at eye level, and their reaction times are recorded.
 - a.** State the null hypothesis.
 - b.** Sketch the appropriate distribution and locate the critical region for the 0.05 level of significance.

c. If the psychologist obtained an average reaction time of $\bar{x} = 195$ milliseconds for this sample, then what decision would be made about the null hypothesis?

d. If the psychologist had used a sample of $n = 100$ subjects and obtained an average reaction time of $\bar{x} = 195$ Milliseconds, then what decision would be made about the effects of the position of the indicator? Explain why this conclusion is different from the one part in part c.

2. A researcher did a one-tailed hypothesis test using an alpha level of 0.01. For this test, H_0 was rejected. A colleague analyzed the same data but used a two tailed test with alpha of 0.05. In this test H_0 was not rejected. Can both analyses be correct? Explain your answer.

3. Suppose a researcher normally uses an alpha level of 0.01 for hypothesis tests but this time used an alpha level of 0.05. What does this change in alpha level do to the amount of power? What does it do to the risk of a Type I error?

4. A researcher wants a statistical tests to be powerful, yet would like to avoid a Type I error. Which of the following approaches would achieve these goals? Explain your answer.

- a.** Increase the alpha level.
- b.** Use smaller alpha but increase the sample size.
- c.** Use a one-tailed test.

5. Patients recovering from an appendix operation normally spend an average of $\mu = 6.3$ days in the hospital. The distribution of recovery time is normal with $\sigma = 1.2$ days. The hospital is trying a new recovery program that is designed to shorten the time patients spend in the hospital. The first 10 appendix patients in this new program were released from the hospital in an average of 5.5 days. On the basis of these data, can the hospital conclude that the new program has a significant effect on recovery time. Test at the 0.05 level of significance.

6. In 1965 a nationwide survey revealed that U.S. grade school children spent an average $\mu = 8.4$ hours per week doing homework. The distribution of homework times was normal with $\sigma = 3.3$. Last year a sample of $n = 200$ students was given the same survey. For this sample, the average number of homework hours was $\bar{x} = 7.1$

a. Do these data indicate a significance change in the amount of homework hours for American grade school children. Test at the 0.01 level of significance.

b. If there had been only $n=20$ students in the sample, would the data still indicate a significance change? Use the same sample mean and α .

7. A family therapist states that parents talk to their teenagers an average of 27 minutes per week. Surprised by that claim, a psychologist decided to collect some data on the amount of time parents spend in conversation with their teenage children. For $n=12$ parents, the study revealed the following time in minutes devoted to conversation in a week: 29, 22, 19, 25, 27, 28, 21, 22, 24, 26, 30, 22 (the mean of this sample is 24.58 and the sample standard deviation is 3.48).

Do the psychologist's findings differ significantly from the therapist's claim? If so, is the family expert's claim an overestimate or underestimate of the actual time spent talking to teenagers? Use 0.05 level of significance with two tail.

8. A group of students complained that all of the statistic classes are offered early in the morning. They claim that they “think better” later in the day and therefore would do better had it been offered in the afternoon. To test this claim, the instructor scheduled a course for 3 p.m. The afternoon class was given the same final the instructor knows that for previous students the scores were normally distributed with a mean of 70. The afternoon class with 16 students had an average score on the final of $\bar{x} = 76$ with $SS=960$. Do the students perform significantly better in the afternoon section? Test with alpha 0.05 and with one-tailed test.

Solution

Q1.

a. $H_0 : \mu = 200$

$$H_1 : \mu \neq 200$$

where μ is the mean of reaction time with the light at eye level.

b. At $\alpha = 0.05$ the z-score = ± 1.96

c. $\bar{x} \approx N(200, \frac{20}{\sqrt{25}}) = N(200, 4)$

c. For

$$\bar{x} = 195,$$

$$z = \frac{\bar{x} - \mathbf{m}}{\mathbf{s}_{\bar{x}}} = \frac{195 - 200}{4} = -1.25$$

So, fail to reject H_0 .

d. $\bar{x} \approx N(200, \frac{20}{\sqrt{100}}) = N(200, 2)$

$$\bar{x} = 195,$$

$$z = \frac{\bar{x} - \mathbf{m}}{\mathbf{s}_{\bar{x}}} = \frac{195 - 200}{2} = -2.5$$

So, reject H_0 .

With larger sample there is less error so the 5 point difference is sufficient to reject H_0 .

Q2.

The analyses are contradictory. The critical region for the two-tailed tests consists of the extremes 2.5% in each tail of the distribution. The two-tailed conclusion indicates that the data was not in the region. However, the critical region for the one-tailed test indicates that the data was on the extreme 1% of one tail. Data cannot be in the extreme 1% and at the same time fail to be in the extreme 2.5%.

Q3.

Increasing the alpha level results in increased power and an increased risk of Type I error.

Q4.

- a.** Increasing alpha would make the test more powerful but it has the undesirable effect of increasing the risk of Type I error.
- b.** Increasing sample size with small alpha would increase power and keep the risk of Type I error small.
- c.** Using a one-tails test would increase power but it would provide an indirect increase in the risk of Type I error.

Q5.

$$H_0 : \mu = 6.3$$

$$H_1 : \mu < 6.3$$

$$\bar{x} \approx N(6.3, \frac{1.2}{\sqrt{10}}) = N(6.3, 0.38)$$

$$\bar{x} = 5.5,$$

$$z = \frac{\bar{x} - \mathbf{m}}{\mathbf{s}_{\bar{x}}} = \frac{5.5 - 6.3}{0.38} = -2.11$$

At $\alpha = 0.05$ the z-score = -1.645

So, reject H_0 (The new program has a significant effect on recovery time.)

Q6. a. $H_0 : \mu = 8.4$ $\bar{x} \approx N(8.4, \frac{3.3}{\sqrt{200}}) = N(8.4, 0.23335)$
 $H_1 : \mu \neq 8.4$

$$\bar{x} = 7.1,$$

$$z = \frac{\bar{x} - \mathbf{m}}{\mathbf{s}_{\bar{x}}} = \frac{7.1 - 8.4}{0.23335} = -5.57$$

At $\alpha = 0.01$ the z-score = ± 2.575

So, reject H_0 (There has been change in homework time.)

b. When $n=20$ $\bar{x} \approx N(8.4, \frac{3.3}{\sqrt{20}}) = N(8.4, 0.7379)$
 $\bar{x} = 7.1,$

$$z = \frac{\bar{x} - \mathbf{m}}{\mathbf{s}_{\bar{x}}} = \frac{7.1 - 8.4}{0.7379} = -1.76$$

So, fail to reject H_0 .

Q7. $H_0 : \mu = 27$

$$H_1 : \mu \neq 27$$

$$\bar{x} \approx t(n - 1) = t(11)$$

$$\bar{x} = 24.58,$$

$$t = \frac{\bar{x} - \mathbf{m}}{s_{\bar{x}}} = \frac{24.58 - 27}{1.01} = -2.40$$

At $\alpha = 0.05$ the t-score = ± 2.201

So, reject H_0 (The data are significantly different “less time” than the therapist;s claim)

Q8.

$$H_0 : \mu = 70$$

$$\bar{x} \approx t(n-1) = t(15)$$

$$H_1 : \mu > 70$$

$$\bar{x} = 76,$$

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{960}{15}} = 8$$

$$t = \frac{\bar{x} - \mathbf{m}}{s_{\bar{x}}} = \frac{76 - 70}{8/\sqrt{16}} = 3$$

At $\alpha = 0.05$ the t-score = 1.753

So, reject H_0 (There is a significance increase in the performance)

Correlation And Regression

Objectives:

- Be able to determine the equation of a simple regression line from a sample data and interpret the slope and intercept of the equation.
- Be able to understand the usefulness of residual analysis in testing the assumptions underlying regression analysis and in examining the fit of the regression line to the data.
- Compute a coefficient of determination and interpret the results.
- Estimate values of Y by using the regression model.
- Compute the coefficient of correlation and interpret it.

In many research situations, the path to decision making lies in understanding the relationships between two or more variables.

Examples:

- A farmer may be interested in the relationship between the level of fertilizer and the yield of tomatoes. In other words the farmer wants to know the level of fertilizer that gives the maximum yield of tomatoes.

- A psychologist may be interested in the relationship between a child's creativity score and flexibility score. More specifically the psychologists would like to know if a child creativity score is a reliable predictor of the child's flexibility score.
- A medical researcher may be interested in the relationship between a patient's blood pressure and the heart rate.

Simple Linear Correlation

To measure the strength of a linear relationship between the two variables one can use the Pearson correlation coefficient **r**. The formula for calculating r is as follows

$$\begin{aligned} r &= \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \\ &= \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \left[\sum y^2 - \frac{(\sum y)^2}{n} \right]}} \end{aligned}$$

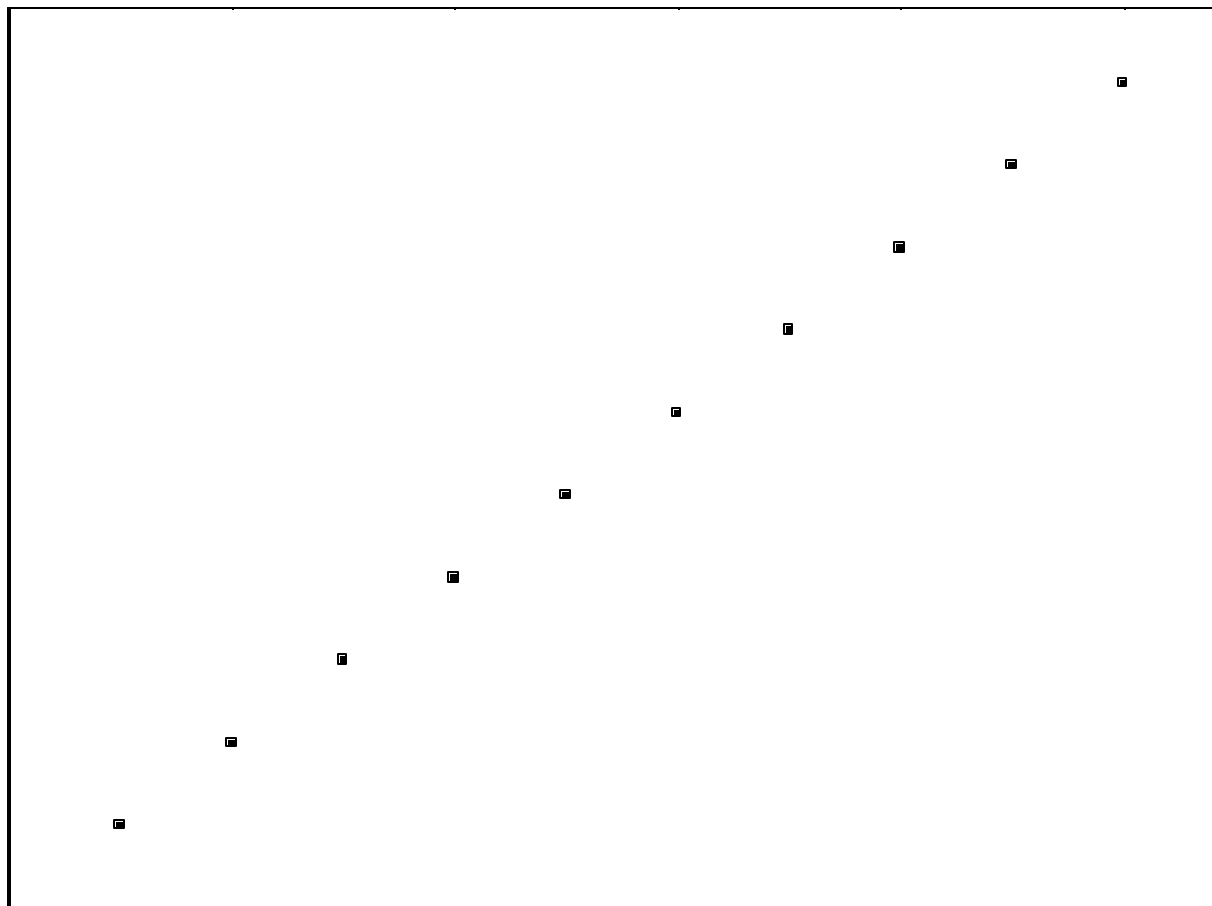
r is a number that ranges between -1 and 1 representing the strength of the relationship.

What are the implications of various possible values of the correlation coefficient r ?

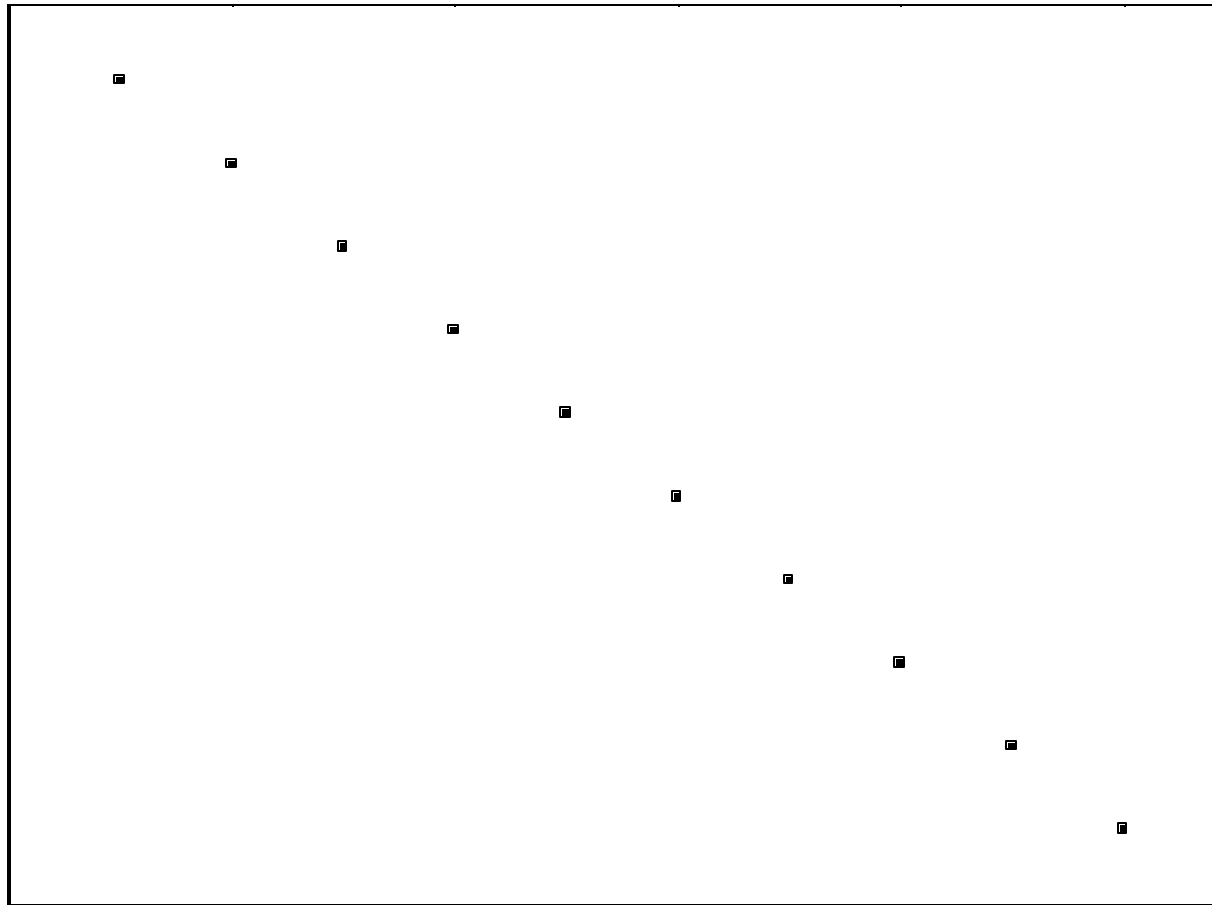
To answer this question let us look at the following scatter plots.

Values of r and their implications

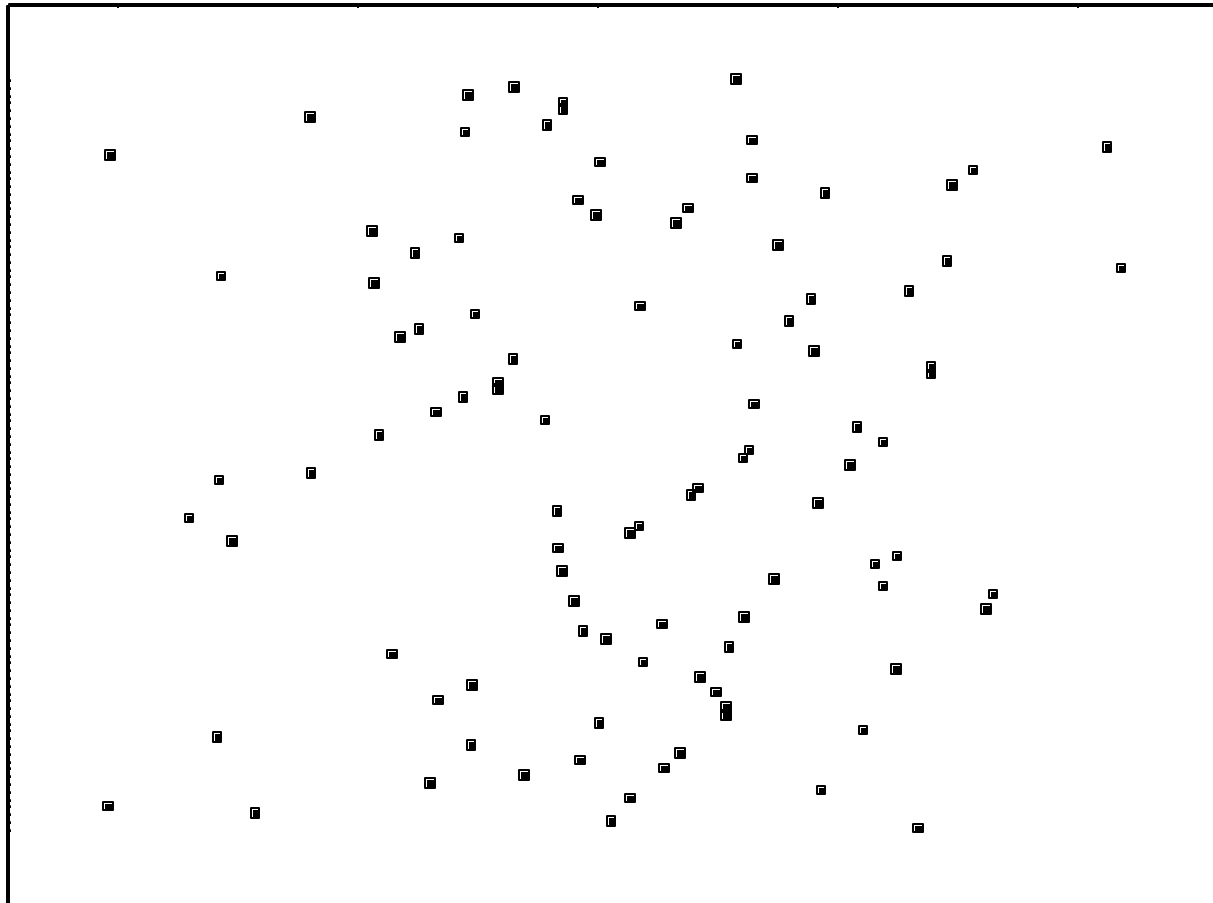
$r = +1$ perfect positive linear relationship between x and y



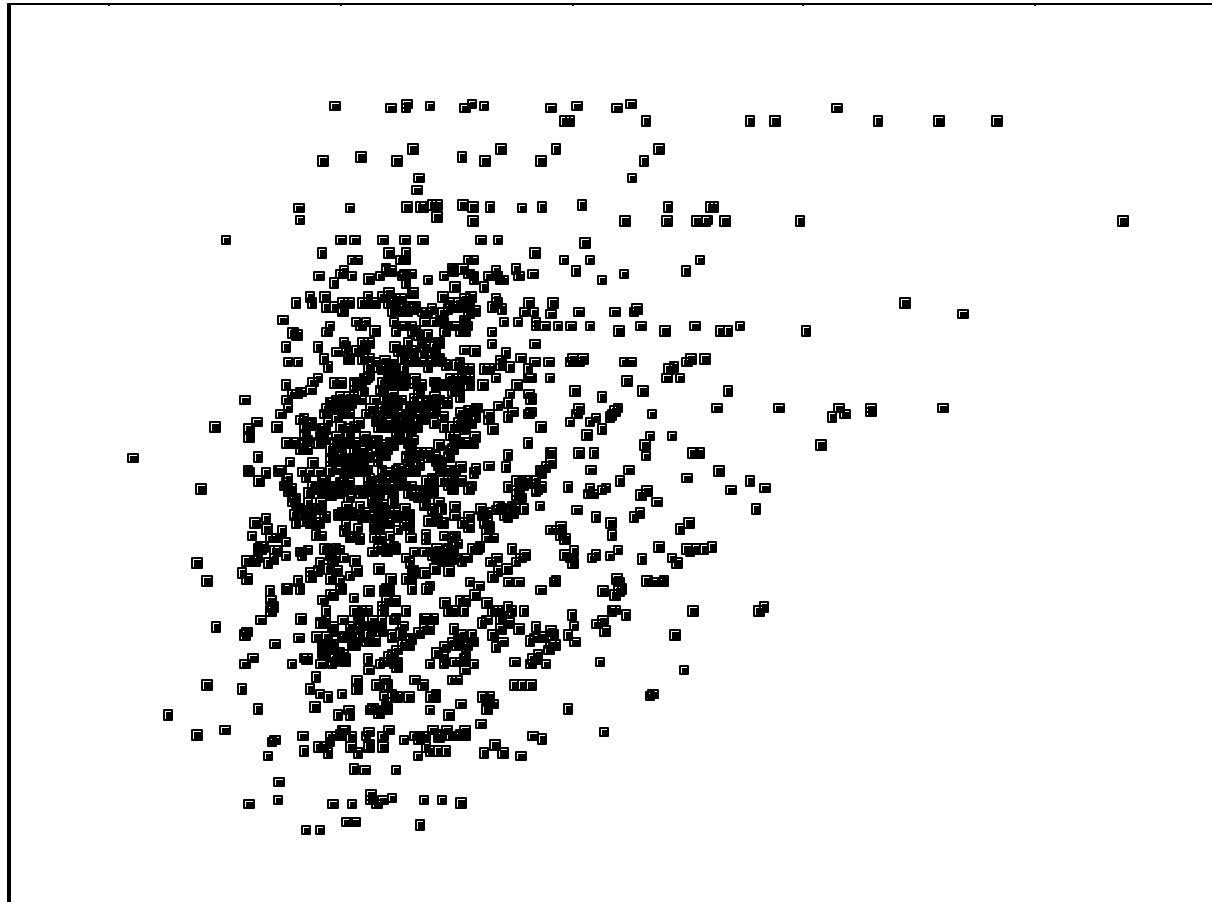
$r = -1$ perfect negative linear relationship
between x and y



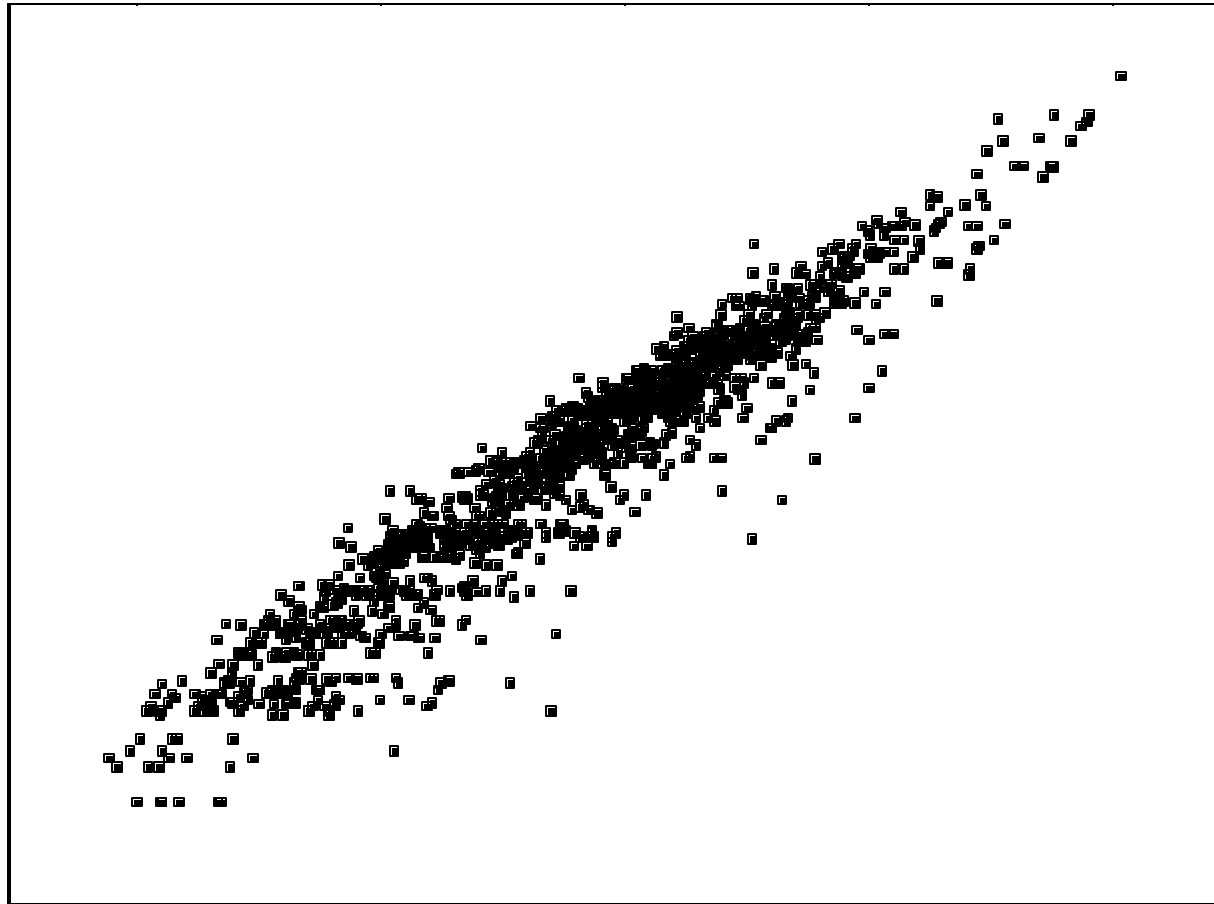
$r = -0.06$ virtually no correlation



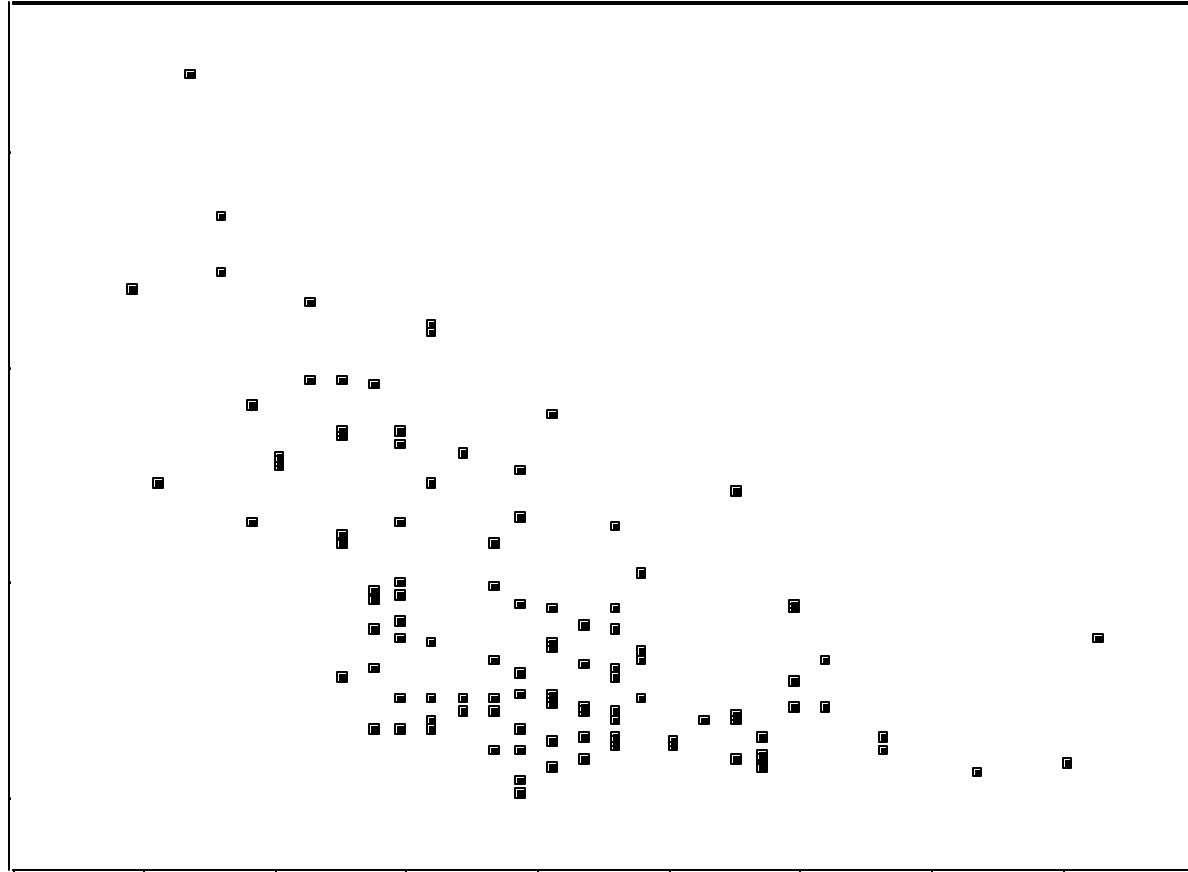
$r = +0.20$ weak positive linear relationship



$r = +0.96$ strong positive linear relationship



$r = -0.60$ moderate negative linear relationship



So,

- An r value of 1 means we have a perfect positive linear relationship.
- An r of -1 means a perfect negative linear relationship.
- An r of 0 means no linear relationship between the two variables.

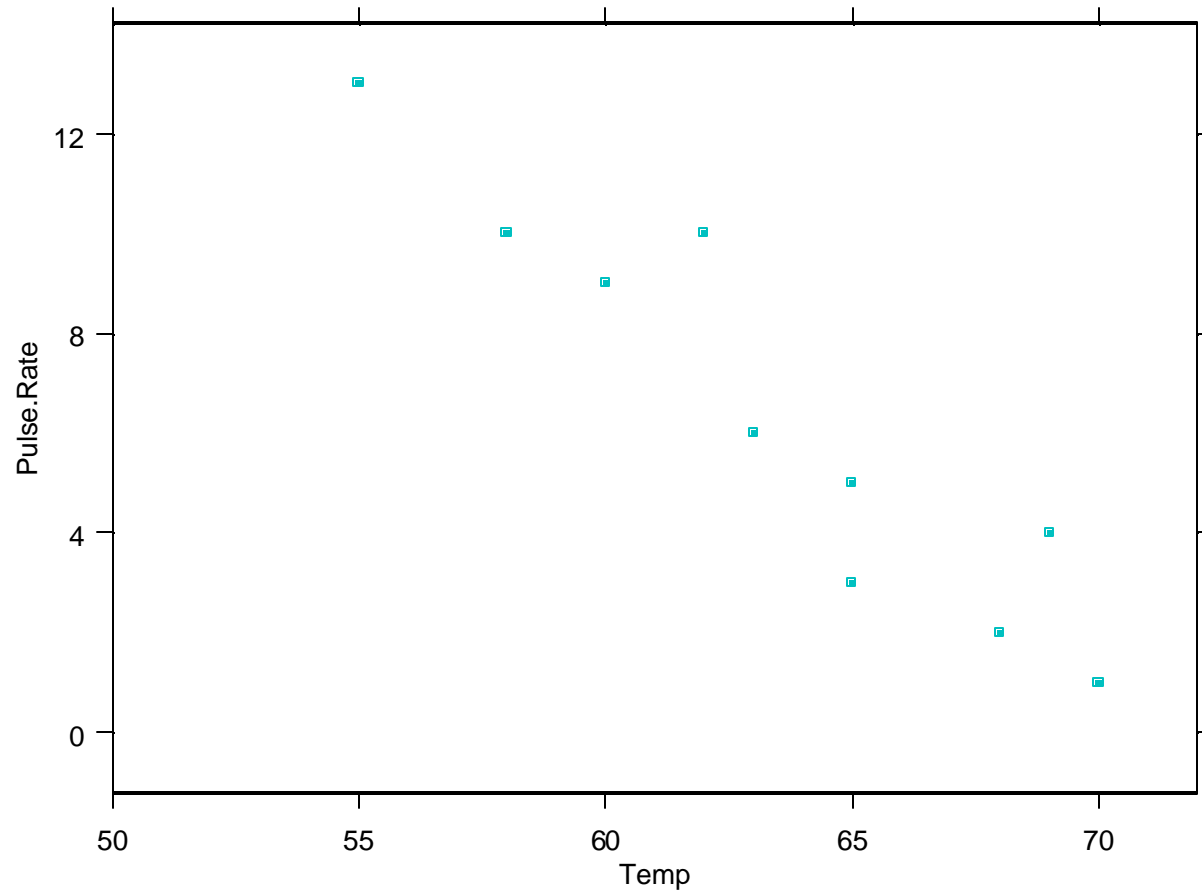
Example 1:

Physicians have used the so-called diving reflex to reduce abnormally rapid heartbeats in humans by submerging the patient's face in cold water. (The reflex triggered by cold water temperatures, is an involuntary neural organs, and sponse that shut off circulation to the skin, muscles, and internal organs, and diverts extra oxygen-carrying blood to the heart, lungs, and brain.) A researcher physician conducted an experiment to investigate the effect of various cold temperatures on the pulse rates of ten small children; the results are presented in the following table.

Table

Child	Temperature of Water x, °F	Reduction in Pulse Rate y, beats/minute
1	68	2
2	65	5
3	70	1
4	62	10
5	60	9
6	55	13
7	58	10
8	65	3
9	69	4
10	63	6

Scatter Plot Of the data



From the scatter plot we can see that the relationship between the temperature of the water and the reduction in pulse rate is strong, linear and negative.

To calculate the value of r we need to use the equation

$$\begin{aligned}
 r &= \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \left[\sum y^2 - \frac{(\sum y)^2}{n} \right]}} \\
 &= \frac{383 - \frac{(635)(63)}{10}}{\sqrt{\left[40537 - \frac{(635)^2}{10} \right] \left[541 - \frac{(63)^2}{10} \right]}} \\
 &= -0.94
 \end{aligned}$$

	x	y	x²	y²	x y
1	6 8	2	4 6 2 4	4	1 3 6
2	6 5	5	4 2 2 5	2 5	3 2 5
3	7 0	1	4 9 0 0	1	7 0
4	6 2	1 0	3 8 4 4	1 0 0	6 2 0
5	6 0	9	3 6 0 0	8 1	5 4 0
6	5 5	1 3	3 0 2 5	1 6 9	7 1 5
7	5 8	1 0	3 3 6 4	1 0 0	5 8 0
8	6 5	3	4 2 2 5	9	1 9 5
9	6 9	4	4 7 6 1	1 6	2 7 6
1 0	6 3	6	3 9 6 9	3 6	3 7 8
T o t a l	6 3 5	6 3	4 0 5 3 7	5 4 1	3 8 3 5

Interpret the value of r for the temperature of water reduction in pulse rate data?

The value of the correlation coefficient $r = -0.94$. So, the implication is that a strong negative relationship between temperature of water and the reduction in pulse rate exist for the ten sampled children. That is, the reduction in pulse rate tends to decrease as the temperature of water increases. However, the research physician should not use this result to conclude that the best way to reduce a child's abnormality rapid heartbeat is to submerge the child's face in extremely cold water.

Since there might be other variables that might contributed to the children reduction of heart rate (for example, the length of time the children are submerged, the children physiological conditions and so on). The only appropriate conclusion to be made is that a negative linear trend may exist between the temperature of water and the reduction in pulse rate.

Understanding And Interpreting The Pearson Correlation

- Correlation simply describes a relationship between two variables. It does not explain why the two variables are related. Specifically, a correlation should not and cannot be interpreted as proof of a cause-and-effect relation between the two variables.
- The value of a correlation can be affected greatly by the range of scores represented in the data.
- One or two extreme data points can have a dramatic effect on the value of a correlation

- When judging “how good” a relationship is, it is tempting to focus on the numerical value of the correlation. For example, a correlation of +0.5 is half way between 0 and 1.00 and therefore appears to represent a moderate degree of relation. However, a correlation should not be interpreted as a proportion. Although a correlation of 1.00 does mean that there is 100% perfectly predictable relation between X and Y, a correlation of 0.5 does not mean that you can make prediction with 50% accuracy. To describe how accurately one variable predicts the other, you must square the correlation. Thus, a correlation of $r = 0.5$ provides $r^2 = 0.25$ or 25% accuracy.

Coefficient of Determination

Definition

The value r^2 is called the coefficient of determination because it measures the proportion of variability in one variable that can be determined from the relationship with the other variable.

Does linear correlation in a sample imply correlation in the population r ? That is, if the calculated value of r is nonzero, does this mean that the population correlation coefficient is nonzero?

The answer is sometimes, but not always. To help us decide we can use Hypothesis testing.

Test of hypothesis for linear correlation

One-Tailed Test

H_0 : There is no linear correlation between the variable x and y (I.e. the population correlation coefficient equals to 0)

H_A : The variables x and y are positively correlated.
(or H_a : The variables x and y are negatively correlated.)

Test statistics: r

Rejection region: $r > r_\alpha$ (or $r < -r_\alpha$)

Two-Tailed Test

H_0 : There is no linear correlation between the variable x and y (I.e. the population correlation coefficient equals to 0)

H_A : The variables x and y are linearly correlated.

Test statistics: r

Rejection region: $r > r_{\alpha/2}$ or $r < -r_{\alpha/2}$

where the distribution of r depends on the sample size n , and r_α and $r_{\alpha/2}$ are the critical values obtained from the following table, such that

$$P(r > r_\alpha) = \alpha \text{ and } P(r > r_{\alpha/2}) = \alpha/2$$

Example2:

Does the data in the previous example provide a sufficient evidence to indicate that the temperature of water and the reduction in pulse rate are linearly correlated. Test using $\alpha = 0.05$.

We need to test the following hypothesis

H_0 : There is no linear correlation between the water temperature and the reduction in the pulse rate.

H_A : There is linear correlation.

Test statistics: $r = -0.94$

$$\alpha = 0.05 \text{ so } \alpha/2 = 0.025$$

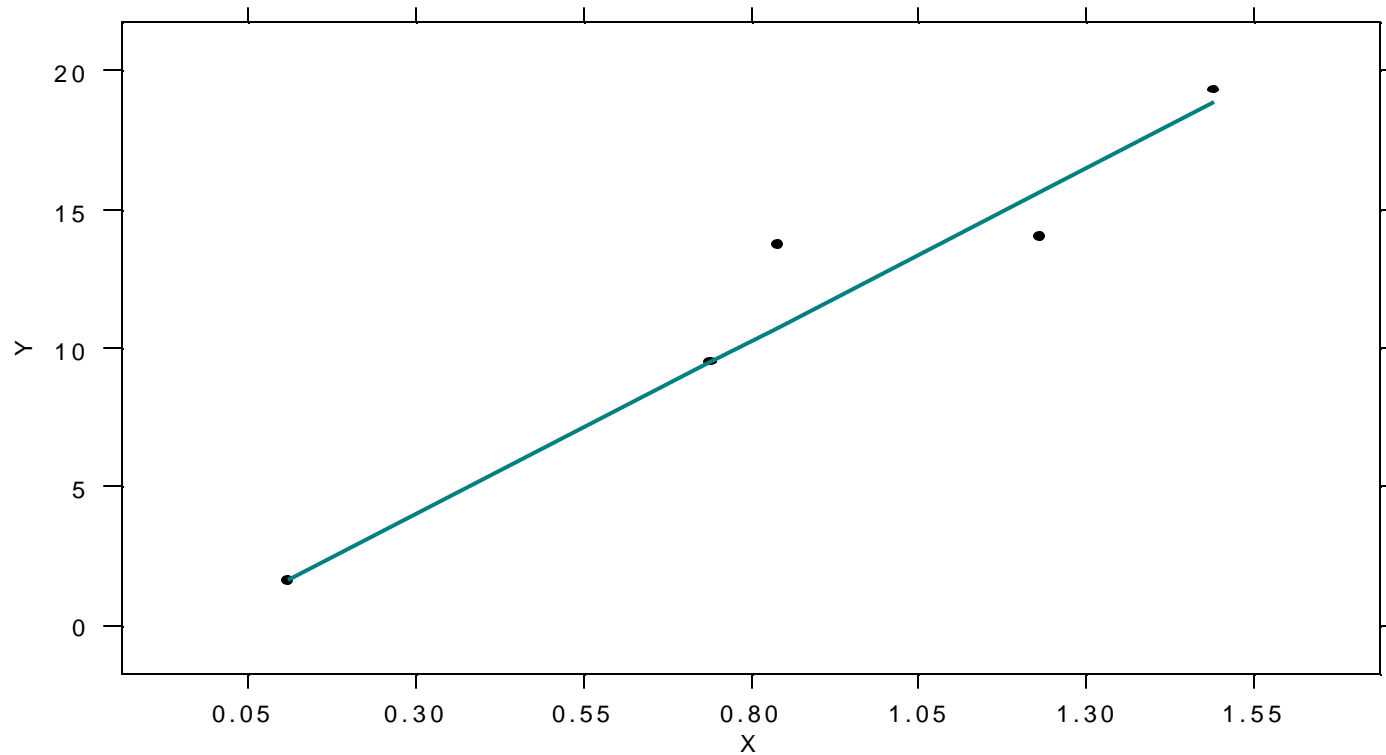
$$r_{\alpha/2} = .632$$

Since $r < r_{\alpha/2}$ we do reject the Null Hypothesis. Which means that there is sufficient evidence that the two variables are correlated.

Simple Linear Regression

The nicotine content and carbon monoxide ranking of a sample of five different cigarette brands given in the following table

Brand	Nicotine Content X, milligrams	CO Ranking Y, milligrams
Camel	.84	13.7
Belair	.74	9.5
Tall	1.49	19.3
Kool	1.23	14.0
Carlton	.11	1.6



The figure shows a good but not perfect, positive relationship. Also note that we have drawn a line through the middle of the data points.

The line serves several purposes:

- The line makes the relationship between nicotine content and carbon monoxide ranking easier to see.
- It provides a simplified description of the relation. So, if the data points were removed, the straight line would still give general picture of the relation between the nicotine content and carbon monoxide ranking.
- The line can be used for predication.

Purpose is to find the line that best fit the data. To do so we are going to use **THE LEAST SQUARE METHOD**. In this method we find the line that minimize the squared distance between the actual Y value and the predicted value (denoted by \hat{Y}) that is determined by the line.

Now,

$$\text{distance} = Y - \hat{Y}$$

this distance is simply the vertical distance between the actual data point (Y) value and the predicted point on the line. This distance measures the error between the line and the actual data. Since some of the values are positive and some are negative. The next step is to square each distance in order to obtain a uniformly positive measure of error. Then sum the squared errors for all of the data points. The result is a measure of overall squared error between the line and the data:

$$\text{total squared error} = \sum (Y - \hat{Y})^2$$

Now, the best fitting line is the one that has the smallest total squared error.

The equation of the line is $\hat{Y} = bX + a$

For each value of X in the data, this equation will determine the point on the line (\hat{Y}) that gives the best prediction of Y .

The problem is to find the values of a and b that makes this line the best fitting line.

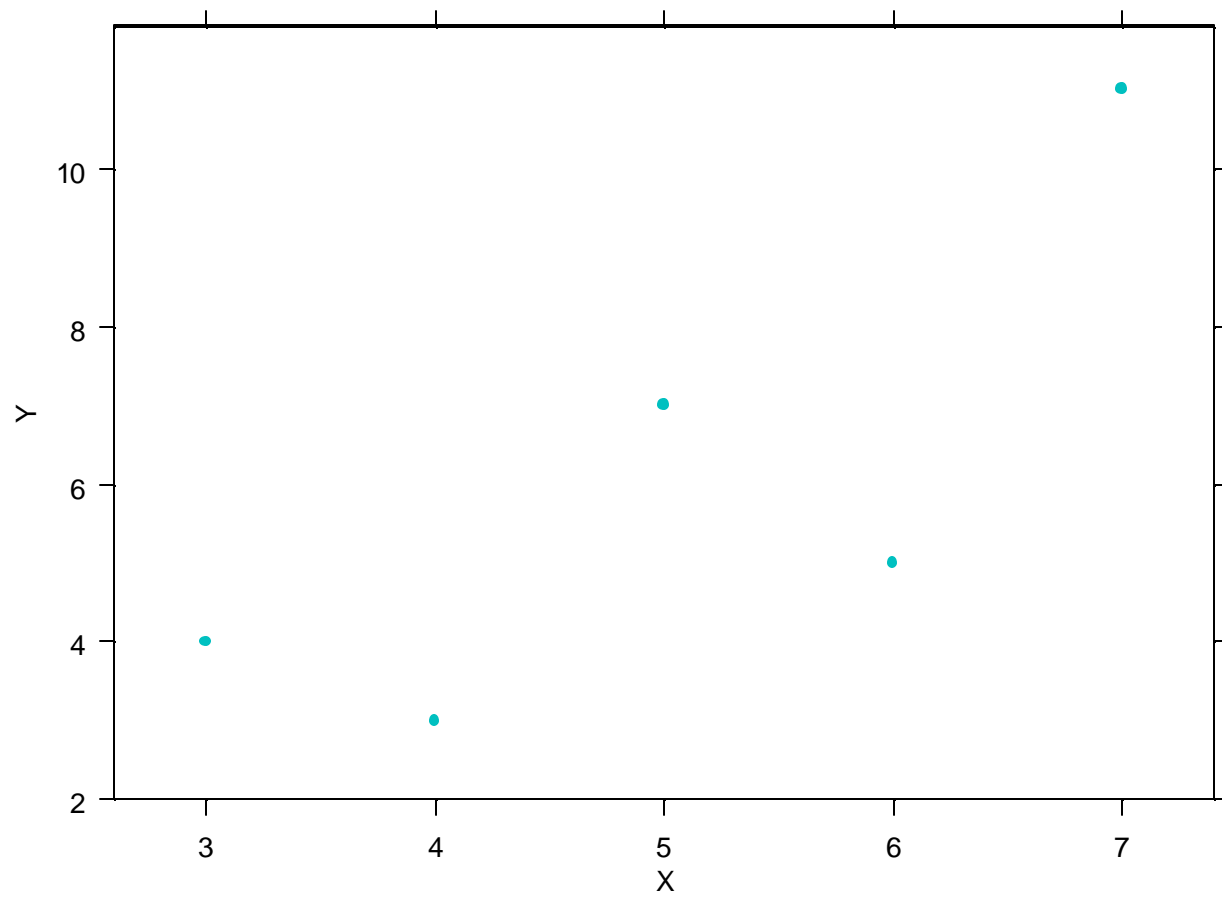
$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

Example 3:

For the following set of data find the linear regression equation for predicting Y from X

X	Y
7	1
4	3
6	5
3	4
5	7



X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
7	11	2	5	10	4
4	3	-1	-3	3	1
6	5	1	-1	-1	1
3	4	-2	-2	4	4
5	7	0	1	0	0

$$\bar{X} = \frac{\sum X}{n} = \frac{25}{5} = 5$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{30}{5} = 6$$

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{16}{10} = 1.6$$

$$a = \bar{Y} - b\bar{X} = -2$$

\Rightarrow

$$\hat{Y} = 1.6X - 2$$

Exercises

Exercises

1. What value does r assume if all the sample points fall on the same straight line and if the line has
 - a. A positive slope?
 - b. A negative slope?

2. The electroencephalogram (EEG) is a device used to measure brain waves. Neurologist have found that the peak EEG frequency in normal children increases with age. IN one study, 287 normal children ranging from 2 to 16 years old were instructed to hold 65-gram weight in the palm of their outstretched hand for a brief but unspecified time. The peak EEG frequency (measured in hertz) was then recorded for each child. A researcher analyzed the data by first grouping the children according to age. He then calculated the average peak frequency for each age group. The data appear in the accompanying table.

Age X , years	Average Peak EEG Frequency Y , hertz	Age X , years	Average Peak EEG Frequency Y , hertz
2	5 . 3 3	1 0	7 . 2 8
3	5 . 7 5	1 1	7 . 0 6
4	5 . 8 0	1 2	7 . 6 0
5	5 . 6 0	1 3	7 . 4 5
6	6 . 0 0	1 4	8 . 2 3
7	5 . 7 8	1 5	8 . 5 0
8	5 . 9 0	1 6	9 . 3 8
9	6 . 2 3		

- A. Construct a scattergram for the data. After examining the scattergram, do you think that x and y are correlated? If correlation is present, is it positive or negative?
- B. Find the correlation coefficient r and interpret its value?
- C. Do the data provide sufficient evidence to indicate that x and y are linearly correlated? Test using $\alpha=0.05$.

3. For the following set of data , find the linear regression equation for predicting Y from X.

X	Y
---	---

0	9
---	---

2	9
---	---

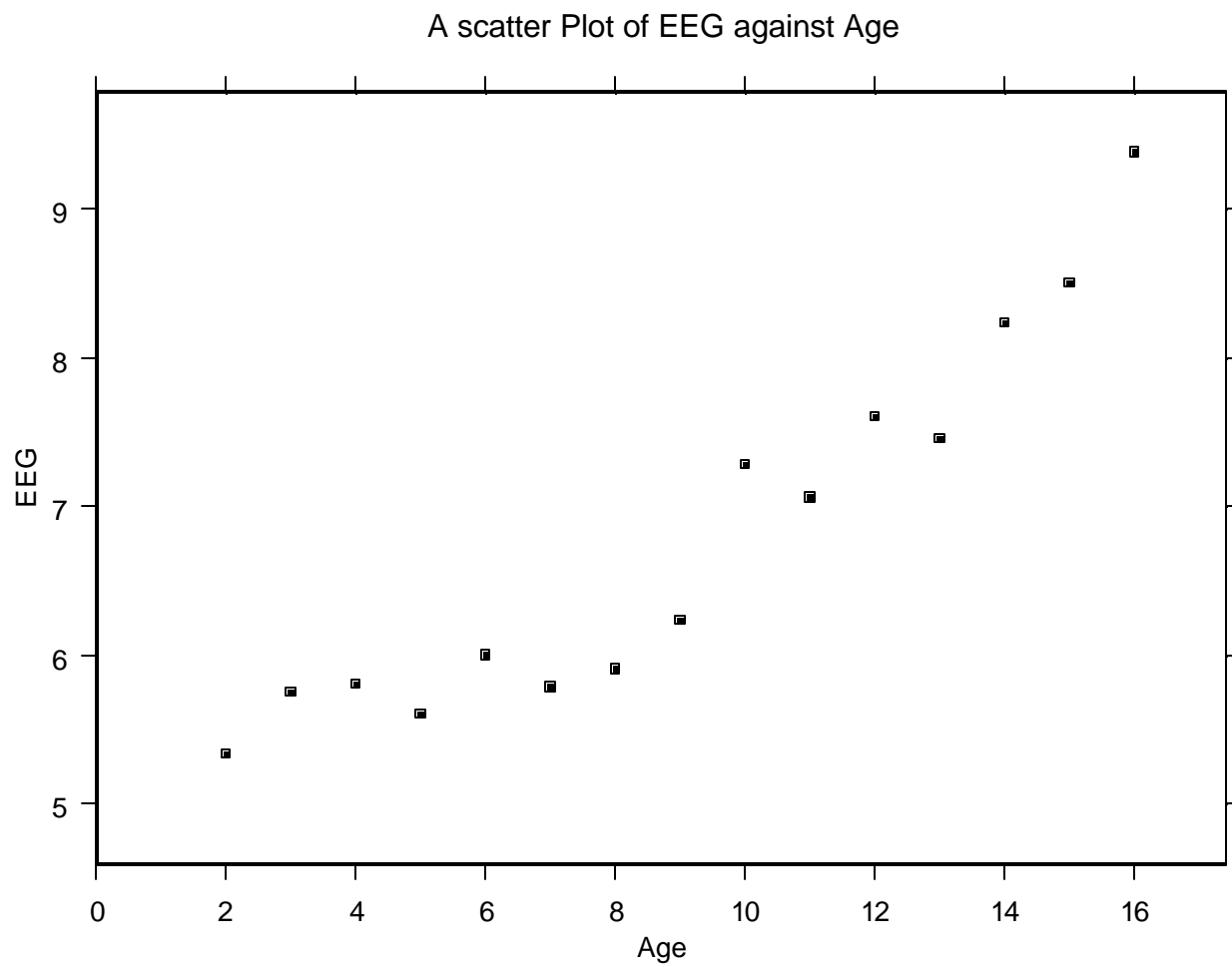
4	7
---	---

6	3
---	---

Solution

- 1.**
 - a.** $r = + 1$
 - b.** $r = - 1$

- 2.**
 - a.** There is a possible positive linear correlation.



b.

$$\begin{aligned} r &= \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \left[\sum y^2 - \frac{(\sum y)^2}{n} \right]}} \\ &= \frac{990.15 - \frac{(135)(101.89)}{15}}{\sqrt{\left[1495 - \frac{(135)^2}{15} \right] \left[713.5745 - \frac{(101.89)^2}{15} \right]}} \\ &= +0.9433 \end{aligned}$$

There is a strong positive linear relationship between age groups and the Average Peak EEG Frequency. That is the frequency tends to increase as the age increases.

c. We need to test the following hypothesis

H_0 : There is no linear correlation between Age and EEG Frequency

H_A : There is linear correlation.

Test statistics: $r = 0.9433$

$$\alpha = 0.05 \text{ so } \alpha/2 = 0.025$$

$$r_{\alpha/2} = .514$$

Since $r > r_{\alpha/2}$ we do reject the Null Hypothesis. Which means that there is sufficient evidence that the two variables are correlated.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
0	9	-3	2	-6	9
2	9	-1	2	-2	1
4	7	1	0	0	1
6	3	3	-4	-12	9

$$\bar{X} = \frac{\sum X}{n} = \frac{0 + 2 + 4 + 6}{4} = \frac{12}{4} = 3$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{9 + 9 + 7 + 3}{4} = \frac{28}{4} = 7$$

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{-20}{20} = -1$$

$$a = \bar{Y} - b\bar{X} = 10$$

\Rightarrow

$$\hat{Y} = 10 - X$$